En tant que chercheurs en sciences sociales, nous sommes fascinés par les questions causales. Dès que nous apprenons que X cause Y, nous voulons mieux comprendre *pourquoi* X cause Y. Ce guide explore le rôle des "mécanismes" dans l'analyse causale et vous aidera à comprendre les types de conclusions que vous pouvez en tirer.

Les mécanismes sont des voies par lesquelles X cause le résultat Y.

Les mécanismes sont depuis longtemps au cœur de la médecine. Chaque fois qu'une médecin prescrit un traitement, elle le fait en sachant quels facteurs chimiques ou physiques provoquent une maladie, et elle prescrit un traitement efficace car il interrompt ces facteurs. Par exemple, de nombreux cliniciens psychologues recommandent l'exercice aux patients souffrant de dépression. L'exercice augmente les endorphines dans la chimie du corps, qui déclenchent des sentiments positifs et agissent également comme analgésiques, ce qui réduit la perception de la douleur. Les endorphines sont donc un mécanisme par lequel l'exercice aide à réduire la dépression. L'exercice peut avoir des effets positifs sur un certain nombre d'autres variables dépendantes (par exemple, les maladies cardiaques) par voie d'autres mécanismes (par exemple, l'élévation du rythme cardiaque), mais le mécanisme qui l'amène à affecter la dépression en particulier est l'endorphine. Nous pourrions également conclure qu'un autre traitement, tel qu'un médicament qui augmente les endorphines, peut avoir des effets similaires sur la dépression.

Les mécanismes sont tout aussi importants pour les sciences sociales. Prenez, par exemple, des recherches récentes qui ont associé le changement climatique à une augmentation des conflits civils. Une étude[1] prétend identifier l'effet causal des chocs climatiques sur les conflits violents en étudiant le taux de conflits civils dans les pays touchés par El Niño pendant les années El Niño par rapport aux années sans El Niño. Supposons que cette étude soit correcte. Pourquoi subir un choc climatique causerait des niveaux de conflit élevés dans un pays ? Un mécanisme pourrait être la pauvreté : les chocs climatiques nuisent à l'économie, et avec des coûts d'opportunité inférieurs, les individus sont plus enclins à rejoindre des groupes armés. Un mécanisme alternatif est physiologique : les gens sont physiquement câblés pour être plus agressifs par temps chaud. Le mécanisme est peut-être la migration : les chocs climatiques déplacent les populations des régions côtières, ce qui produit des conflits sociaux entre migrants et autochtones. En réalité, plusieurs ou tous ces mécanismes (ainsi que d'autres non listés ici) pourraient fonctionner simultanément ! Dans bon nombre des questions les plus intéressantes en sciences sociales, il existe plusieurs canaux (ou mécanismes "M") qui pourraient transmettre l'effet total de X sur Y.

Bien que nous n'ayons pas *besoin* de connaître le mécanisme pour conclure que X cause Y, il y a plusieurs raisons pour lesquelles nous le *voudrions*.

Dans l'exemple climat/conflit ci-dessus, nous pouvons avoir une confiance totale dans la capacité des chercheurs à identifier que les chocs climatiques causent des conflits, et pourtant n'avons aucune preuve des mécanismes qui sont à l'œuvre. Mais les chercheurs en sciences sociales aiment en apprendre davantage sur les mécanismes parce qu'ils sont étroitement liés aux théories des sciences sociales. Par exemple, le mécanisme de "pauvreté" ci-dessus est étroitement lié à la théorie de Gurr[2] selon laquelle les individus se rebellent lorsque les coûts d'opportunité du conflit sont faibles, alors que le mécanisme de "migration" pourrait soutenir une théorie du conflit basée sur des griefs entre groupes sociaux. Il n'est pas étonnant qu'en apprenant que X cause Y, les chercheurs en sciences sociales se demandent immédiatement quel est le mécanisme - ils veulent relier cette découverte à la théorie !

[1] Solomon M. Hsiang, Kyle C. Meng et Mark A. Cane, "Civil Conflicts Are Associated with the Global Climate," Nature 476.7361 (2011): 438-441.

[2] Ted Gurr, Why Men Rebel, Princeton University Press, 1970.

Comprendre les mécanismes a non seulement des avantages théoriques mais aussi pratiques. Premièrement, connaître M permet de deviner pour quelles populations X conduira à Y. Si le mécanisme du climat/conflit est une réponse physiologique à la chaleur, alors les chocs climatiques peuvent produire des conflits uniquement lorsque la température est assez élevée. Deuxièmement, connaître M nous aide à considérer d'autres résultats qui peuvent être affectés par X. Si le mécanisme du climat/conflit est la migration, alors nous pouvons également nous attendre à ce que les chocs climatiques entraînent une surutilisation des biens publics dans les zones urbaines. Troisièmement, connaître M nous aide à envisager d'autres moyens de provoquer ou d'éviter de provoquer des changements dans Y. Si le mécanisme du climat/conflit est la pauvreté, alors les programmes de développement pourraient diminuer les conflits en réduisant la sensibilité des revenus aux chocs climatiques, même s'ils ne peuvent pas changer les chocs climatiques.

Mais il est extrêmement difficile d'identifier les mécanismes causaux car les mécanismes eux-mêmes ne sont pas assignés de manière aléatoire. . .

Prenons un exemple expérimental. Chong et al. (2015)[3] ont utilisé une expérience de terrain pour étudier l'effet des informations à propos de la corruption sur la participation électorale. Ils ont assigné de manière aléatoire certains bureaux de vote au Mexique pour recevoir des informations sur l'utilisation corrompue des fonds au sein de cette municipalité. Étonnamment, ils ont découvert que les circonscriptions traitées votaient à des taux inférieurs aux circonscriptions de contrôle. Ils suggèrent le mécanisme suivant à l'œuvre : les informations sur la corruption convainquent les électeurs que la municipalité est si gravement corrompue que l'élection d'un bon politicien ne la changera pas, de sorte que les individus trouvent que leur vote a moins de valeur.

En bref, leur argument est :[4]

Réception d'informations sur la corruption (X) $\xrightarrow{+}$ Estime que la corruption est trop grave (M) $\xrightarrow{+}$ Reste à la maison (Y)

Chong et al. font face à un obstacle commun dans l'interprétation de leurs résultats : le mécanisme qu'ils proposent n'a pas été assigné de manière aléatoire. Certaines personnes sont plus enclines à croire que "tous les politiciens sont des vauriens", tandis que d'autres ont tendance à faire pression pour "un changement auquel nous pouvons croire". Malheureusement, nous ne pouvons observer que le traitement aléatoire qu'un individu a reçu et sa croyance (non aléatoire) au sujet de la corruption ; nous ne pouvons pas dire quelle croyance au sujet de la corruption ils *auraient eu* s'ils avaient reçu l'autre condition de traitement. Il nous est donc impossible de déterminer, pour chaque individu, dans quelle mesure sa décision d'aller voter a été causée par le mécanisme proposé par rapport à d'autres mécanismes.

Certains chercheurs tentent de contourner ce problème en estimant l'effet *moyen* du traitement sur le mécanisme, puis en estimant l'effet *moyen* du mécanisme sur le résultat. L'une des raisons pour lesquelles cela est problématique est que nous pouvons imaginer plusieurs facteurs autres que le traitement qui pourraient causer à la fois M et Y. Supposons que le niveau d'apathie - appelons-le Q - varie parmi les citoyens de notre étude, et Q a un effet très fort sur M et Y. Les personnes très apathiques pourraient être plus susceptibles de croire que les problèmes sont insolubles, et elles pourraient également être plus susceptibles de rester à la maison le jour du scrutin. Nous sommes donc susceptibles d'observer une forte corrélation entre M et Y qui est induite par le facteur de confusion Q, et non par notre traitement X. Mécaniquement, nos résultats seront biaisés en faveur de la recherche de preuves de l'effet de X sur Y via M simplement parce que Q a produit une relation entre M et Y.

[3] Alberto Chong, Ana L. de la O, Dean Karlan et Leonard Wantchekon, "Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification," The Journal of Politics 77.1 (2015): 55-71.

[4] Notez que Chong et al. testent leur argument en utilisant des données au niveau de l'arrondissement, pas au niveau individuel, mais nous avons adapté leur argument au niveau individuel pour faciliter l'explication.

. . . et parce que l'effet du traitement est rarement homogène.

L'autre problème lorsque l'on essaie de décomposer l'effet moyen de X sur M puis de M sur Y est que cette approche suppose que chaque sujet réponde au traitement de manière identique. En revenant à notre exemple dans lequel X est le traitement de l'information, M est la conviction que la corruption est trop sévère, et Y le fait de rester à la maison, nous pouvons imaginer deux types de répondants. Le type A pensait que la corruption était trop grave pour être résolue jusqu'à ce qu'elle reçoive une lettre contenant des informations sur la corruption dans son district. Elle fut surprise de voir que le problème n'était pas aussi grave qu'elle l'avait pensé. Formellement, pour le type A, $M(X = 0) = 1$ et $M(X = 1) = 0$, donc X a un effet *négatif* sur M. Le type B pensait que la corruption était un problème gérable jusqu'à ce qu'elle reçoive une lettre contenant des informations sur la corruption dans son district. Elle a été surprise par l'ampleur du problème et a abandonné tout espoir de résoudre le problème. Formellement, pour le type B, $M(X = 0) = 0$ et $M(X = 1) = 1$, donc X a un effet *positif* sur M. Si nous devions faire la moyenne des effets pour ces deux types, nous ne verrions aucune relation entre X et M.

Type	X (traitement de l'info)	M conditionnel à X=0 (non observé)	M conditionnel à X=1 (observé)	Effet de X sur M	Effet de M sur Y	Y (Reste à la maison)
A	1	1	0	négatif	négatif	1
B	1	0	1	positif	positif	1

L'estimation du rôle de M peut être encore plus compliquée lorsque la relation entre M et Y est également hétérogène. Imaginez que le type A ne vote que lorsqu'elle est en colère (en d'autres termes, M a un effet *négatif* sur Y). Le type A prévoyait de voter pour exprimer sa colère face à l'omniprésence de la corruption dans son district, même si elle savait que cela n'aurait rien changé, jusqu'à ce qu'elle apprenne que la corruption n'était pas aussi grave qu'elle l'avait imaginée. Sa colère disparue, elle choisit de rester à la maison le jour des élections. Cependant, le type B ne vote que lorsqu'elle pense que son vote peut faire la différence (en d'autres termes, M a un effet *positif* sur Y). Le type B allait voter pour les politiciens non corrompus de son district jusqu'à ce qu'elle apprenne qu'ils étaient tous corrompus. Sans aucun espoir de changer la situation, elle a également décidé de rester chez elle le jour du scrutin. Pour les types A et B, il existe un "effet indirect" de M (en d'autres termes, X affecte Y à travers M). Mais nous passerons à côté de cette relation dans l'ensemble car nous ne serons pas en mesure d'obtenir des estimations non biaisées de l'effet moyen de X sur M.[5]

Nous pouvons imaginer beaucoup plus de "types" que A et B – le but ici est de démontrer intuitivement que parce que M n'est pas assigné de manière aléatoire, et parce qu'il est peu probable que les effets de X sur M *et* M sur Y soient identiques pour tout le monde, il sera très difficile de caractériser avec précision l'entremise de M pour notre effet.

De nombreuses études tentent de décomposer l'effet total d'un traitement entre ses effets "directs" et "indirects".

Parce que l'apprentissage des mécanismes est riche en promesses théoriques, les chercheurs aimeraient quantifier dans quelle mesure un effet de X sur Y opère via M. Parfois, les chercheurs essaieront de le faire grâce à une technique appelée "décomposition des effets".

Une analyse de décomposition des effets tente de décomposer un effet *total* de X sur Y en l'effet que X a sur Y *directement* et l'effet de X sur Y qui se produit *indirectement* via M. "L'effet total" fait référence

[5]Pour une discussion plus rigoureuse de ces fausses conceptions, voir Adam N. Glynn, "The Product and Difference Fallacies for Indirect Effects," American Journal of Political Science 56.1 (2012): 257-269.

à l'effet moyen du traitement (average treatment effect, ATE), qui est simplement l'effet moyen de X sur Y. Toute expérience qui assigne de manière aléatoire un traitement afin d'observer ses effets sur certains résultats évalue l'ATE. Ensuite, le chercheur essaie de quantifier la taille de l'effet que X a sur Y à travers le mécanisme M. C'est ce qu'on appelle souvent "l'effet indirect" – parce que X affecte indirectement Y via M – ou l'effet moyen causal par médiation (average causally mediated effect, ACME). Enfin, le chercheur tentera d'estimer l'effet de X sur Y qui ne passe pas par M. C'est ce qu'on appelle "l'effet direct" de X sur Y ou l'effet moyen direct contrôlé (average controlled direct effect, ACDE), car c'est l'effet de X sur Y lorsque nous contrôlons le travail effectué par M.

Mais soyez prudent lorsque vous utilisez une régression pour décomposer les effets.

Bien que couramment utilisée, l'analyse de régression par médiation présuppose des hypothèses fortes et souvent irréalistes. Nous utiliserons du code pour illustrer ce que cette méthode implique et démontrer les conditions dans lesquelles elle peut produire des estimations biaisées.

L'idée de base est que si nous avons des données sur le traitement qu'un individu a reçu (X), s'il présente le mécanisme proposé (M) et quel est le résultat (Y), alors nous pouvons distinguer ces effets en utilisant les trois régressions suivantes.

(1)
$$M_i = \alpha_1 + aX_i + e_{1i}$$

(2)
$$Y_i = \alpha_2 + cX_i + e_{2i}$$

(3)
$$Y_i = \alpha_3 + dX_i + bM_i + e_{3i}$$

Comment ferions-nous cela? En utilisant l'équation 1, nous faisons la régression de M sur X pour obtenir l'effet direct de X sur M, qui est le coefficient a. Ensuite, nous passons à l'équation 3, dans laquelle nous faisons la régression de Y sur M et X. Dans cette régression, le coefficient b représente l'effet direct de M sur Y lorsque nous contrôlons pour X. Une analyse de décomposition des effets multiplierait $a*b$ pour révéler l'effet indirect de X sur Y via M. Pour trouver l'effet direct de X sur Y, nous n'avons pas besoin de chercher plus loin que d, qui est le coefficient de X dans l'équation 3 lorsque nous contrôlons pour M. En d'autres termes, d est l'effet de X sur Y qui ne passe pas par M. Si l'on additionne l'effet indirect et l'effet direct, on obtient "l'effet total" de X sur Y, qui est égal à c. Pour résumer, l'analyse de décomposition des effets désagrège ostensiblement l'effet total en effet qui est médié via M et l'effet qui n'est pas médié via M, permettant au chercheur de conclure à quel point M est important pour expliquer la relation entre X et Y.[6]

Le problème est que cette arithmétique ne fonctionne que sous certaines hypothèses très fortes. L'une de ces hypothèses est que les termes d'erreur dans les régressions 1 et 3 ne sont pas liés les uns aux autres - en d'autres termes, M ne peut pas être prédit par des facteurs non observables qui prédisent également Y. Nous avons décrit ce problème intuitivement au point 3 lorsque nous avons introduit Q, une variable de confusion qui contribue à la fois à M et à Y et engendre donc une relation très forte entre eux, même si l'effet de X sur Y ne s'exerce pas du tout à travers M. Décrivons maintenant ce problème à l'aide d'une simulation.

Dans le code suivant, nous commençons par créer cette variable Q pour chaque individu et définir les "vrais" effets de X sur M, M sur Y et X sur Y. Ensuite, nous créons des résultats potentiels hypothétiques pour M, c'est-à-dire que pour chaque individu, nous définissons quelle valeur de M ils révéleraient s'ils étaient traités, et quelle valeur de M ils révéleraient s'ils n'étaient pas traités. Ces valeurs sont liées non seulement au "vrai" effet de X sur M, mais aussi à Q. Ensuite, nous pouvons également définir des résultats potentiels

[6] Explication adaptée d'Alan Gerber et Donald Green, Field Experiments, W.W. Norton and Company, 2012, chapter 10.

hypothétiques pour Y. Nous faisons cela pour quatre scénarios, qui supposent tous des effets constants, une hypothèse que nous relâcherons plus tard. Deux d'entre eux sont des résultats potentiels simples de Y : le Y présenté par l'individu qui n'est pas traité et révèle son résultat potentiel M non traité, et le Y présenté par l'individu qui est traité et révèle son résultat potentiel M traité. Cependant, nous définissons également deux résultats potentiels complexes de Y : le Y présenté par l'individu qui n'est pas traité mais qui révèle son résultat potentiel M traité, et le Y présenté par l'individu qui est traité mais qui révèle son résultat potentiel M non traité. Bien que ces résultats potentiels compliquent la compréhension, ils sont importants à définir afin que nous puissions calculer les effets directs et indirects "vrais" (mais intrinsèquement inobservables) auxquels comparer notre analyse de décomposition.

Dans la seconde moitié du code, nous effectuons une assignation aléatoire du traitement et procédons à l'analyse de décomposition des effets décrite ci-dessus, en utilisant les données que nous "observons". Sous les hypothèses (fortes) selon lesquelles les termes d'erreur ne sont pas corrélés et les effets sont constants entre les sujets, $a * b =$ ACME, $d =$ ACDE et $c =$ ATE. Cependant, la simulation révèle que $a * b >$ ACME et $d <$ ACDE ; c'est-à-dire que nous avons surestimé l'ACME et sous-estimé l'ACDE. Notre analyse de décomposition des effets était biaisée car la première hypothèse - termes d'erreur non corrélés - n'était pas vérifiée : la variable non observée Q prédisait à la fois M et Y, ce qui nous a conduit à surestimer le rôle du mécanisme M.

```
rm(list = ls())

set.seed(20160301)

N <- 1000000

# simuler des données, créer des résultats potentiels (RP), estimer les effets "réels"

# construire une caractéristique non observée idéosyncratique
Q_i <- rnorm(N)

# créer le "vrai modèle" en définissant l'effet du traitement (tau)
tau_X_on_M <- 0.2 # effet de X sur M
tau_M_on_Y <- 0.1 # effet de M sur Y
tau_X_on_Y <- 0.5 # effet total de X sur Y (ATE), à travers M et non M

# construire les résultats potentiels pour le médiateur
# individu révèle M_1 s'il est traité ; M_0 si non traité
# M est fonction à la fois du traitement et de la caractéristique non observée
M_0 <- 0 * tau_X_on_M + Q_i
M_1 <- 1 * tau_X_on_M + Q_i

# nous pouvons estimer l'effet moyen du traitement (ATE) non biaisé de X sur M
ATE_M <- mean(M_1 - M_0)
ATE_M
```

```
[1] 0.2
```

```
# construire les résultats potentiels (RP) pour le résultat
Y_M0_X0 <- tau_M_on_Y * (M_0) + tau_X_on_Y * 0 + Q_i
Y_M1_X1 <- tau_M_on_Y * (M_1) + tau_X_on_Y * 1 + Q_i
Y_M0_X1 <- tau_M_on_Y * (M_0) + tau_X_on_Y * 1 + Q_i # Ceci est un RP "complexe"
Y_M1_X0 <- tau_M_on_Y * (M_1) + tau_X_on_Y * 0 + Q_i # Ceci est un RP "complexe"

# certains de ces RP sont "complexes" car nous imaginons ce que nous
```

```r
# observerions si nous avions assigné au traitement mais observé le RP de M non traité ou
# si nous avions assigné au contrôle mais observé le RP de M traité
# la construction de ces RP complexes est nécessaire pour estimer les "vrais" effets directs et indirec

# nous pouvons estimer l'ACME
# nous estimons les effets de M en maintenant X constant
# ce sont les mêmes
# c'est "l'effet indirect"
ACME_X0 <- mean(Y_M1_X0 - Y_M0_X0)
ACME_X1 <- mean(Y_M1_X1 - Y_M0_X1)
ACME <- mean(((Y_M1_X1 - Y_M0_X1) + (Y_M1_X0 - Y_M0_X0)) / 2)

# nous pouvons estimer l'ACDE
# nous estimons les effets de X en maintenant M constant
# ce sont les mêmes
# c'est "l'effet direct"
ACDE_M0 <- mean(Y_M0_X1 - Y_M0_X0)
ACDE_M1 <- mean(Y_M1_X1 - Y_M1_X0)
ACDE <- mean(((Y_M0_X1 - Y_M0_X0) + (Y_M1_X1 - Y_M1_X0)) / 2)

# maintenant, nous construisons les RP simples pour Y
Y_1 <- tau_M_on_Y * (M_1) + tau_X_on_Y * 1 + Q_i
Y_0 <- tau_M_on_Y * (M_0) + tau_X_on_Y * 0 + Q_i

# on estime le vrai ATE de X sur Y
# c'est l'effet "total"
ATE <- mean(Y_1 - Y_0)

ATE
```

```
[1] 0.52
```

```r
ACDE + ACME # noter que les effets directs et indirects s'additionnent au total
```

```
[1] 0.52
```

```r
ACDE
```

```
[1] 0.5
```

```r
ACME
```

```
[1] 0.02
```

```r
ATE_M
```

```
[1] 0.2
```

```r
# Assignation aléatoire, révélation des RPs, tentative de décomposition des effets

# nous assignons la moitié de notre échantillon au traitement et l'autre moitié au contrôle
X <- sample(c(rep(1, (N / 2)), rep(0, (N / 2))))
# nous révélons les RPs pour M et Y en fonction de l'assignation de traitement
M <- X * M_1 + (1 - X) * M_0
Y <- X * Y_1 + (1 - X) * Y_0

model1 <- lm(M ~ X)
a <- coef(model1)[2] # extrait le coefficient pour obtenir l'effet de X sur M
a
```

```
  X
0.2
```

```r
model2 <- lm(Y ~ X)
c <- coef(model2)[2] # extrait le coefficient pour obtenir l'effet total de X sur Y
c
```

```
   X
0.52
```

```r
model3 <- lm(Y ~ X + M)
d <- coef(model3)[2] # extrait le coefficient pour obtenir l'effet de X sur Y en contrôlant pour M
b <- coef(model3)[3] # extrait le coefficient pour obtenir l'effet de M sur Y en contrôlant pour X

# on pourrait multiplier maintenant l'effet moyen de X sur M et l'effet moyen de M sur Y
# pour obtenir l'ACME de X sur Y via M
a * b
```

```
   X
0.22
```

```r
# mais quand on compare cela au vrai ACME, on voit que c'est biaisé
ACME
```

```
[1] 0.02
```

```r
# on pourrait également interpréter l'effet moyen de X sur Y en contrôlant M comme l'ACDE
d
```

```
  X
0.3
```

```r
# mais quand on compare cela au vrai ACDE, on voit que c'est biaisé
ACDE
```

```
[1] 0.5
```

```
# notez que nous avons SUR-estimé l'effet indirect moyen et SOUS-estimé l'effet direct moyen

# les estimations qui ne sont pas biaisées sont les effets moyens de X sur Y et de X sur M
# car X est assigné de manière aléatoire
a
```

```
  X
0.2
```

```
ATE_M
```

```
[1] 0.2
```

```
c
```

```
   X
0.52
```

```
ATE
```

```
[1] 0.52
```

Relions cet exercice à la question soulevée au point 3. Cette simulation a illustré que la quantification de l'effet de médiation s'avère difficile lorsque les variables prédictives de base confondent la relation entre M et Y. Étant donné que M n'est pas assigné de manière aléatoire, il est important pour nous de réfléchir à la probabilité que notre M et notre Y soient tous deux affectés par des variables non observées. En principe, s'il n'y a pas de variables de confusion dans cette relation, alors une analyse de décomposition des effets peut être non biaisée, mais cette hypothèse est forte et généralement difficile à prouver.

Même si nous ne l'avons pas démontré dans cette simulation, il est également possible de montrer que la décomposition des effets est également inappropriée lorsque l'effet du traitement est hétérogène (nous avons introduit cette intuition au point 4). La raison technique de ceci vient de notre loi de l'espérance, qui est : $E[a * b] = E[a]E[b] + cov(a, b)$. Si nous avons un effet de traitement constant, alors a et b ne covarient pas, le terme de covariance disparaît et nous pouvons simplement multiplier $a * b$ pour obtenir l'ACME. Cependant, si le terme de covariance est non nul, alors nous ne sommes pas en mesure d'estimer cet effet indirect à partir de ces deux coefficients obtenus à partir de régressions distinctes. Nous avons construit un effet de traitement constant afin de pouvoir démontrer le processus d'effets de décomposition, mais si nous devions refaire la simulation avec un effet de traitement hétérogène qui varie, nous ne serions même pas en mesure de calculer l'ACME ou l'ACDE en utilisant l'approche des résultats potentiels au début du code.

Ce que vous pouvez faire... Avant de vous lancer dans une analyse de décomposition des effets, demandez-vous :

- Puis-je imaginer des variables non observées qui prédisent à la fois M et Y ?
- Est-il possible que mes sujets réagissent à l'effet du traitement de différentes manières ?

Si la réponse à l'une de ces questions est oui, nous vous recommandons fortement de procéder avec prudence. En particulier, réfléchissez bien à la manière dont les variables non observées et l'effet du traitement hétérogène affecteraient votre stratégie d'estimation.

Parfois, l'analyse de sous-groupes peut fournir des preuves ou suggestions pour ou contre un mécanisme.

Aux points 3-6, nous avons mis en garde les chercheurs contre toute tentative de quantifier avec confiance la proportion d'un effet médié par un mécanisme particulier, mais il peut exister d'autres moyens d'en savoir plus sur les mécanismes à l'œuvre dans une étude particulière. Au point 1, nous avons souligné la relation étroite entre les mécanismes et la théorie. Ce n'est pas parce qu'il est difficile de quantifier directement les preuves d'un mécanisme que nous ne pouvons pas explorer les prédictions testables de la théorie dans laquelle notre mécanisme est présenté !

Une stratégie consiste à utiliser l'analyse de sous-groupe, ou des interactions de traitement par covariable, pour voir si différentes populations répondent au traitement différemment conformément à nos théories. Par exemple, supposons que nous voulions en savoir plus sur le rôle du revenu dans la médiation de la relation climat/conflit. L'une des implications vérifiables d'une théorie dans laquelle le revenu joue un rôle médiateur est que nous nous attendrions à ce que les chocs climatiques soient associés à des conflits dans les zones où le revenu est sensible aux chocs climatiques mais pas là où le revenu est indépendant des chocs climatiques. Sarsons (2015)[7] fait exactement cela. Exploitant le fait que les districts en aval des barrages d'irrigation ne dépendent pas des précipitations pour leur revenu contrairement aux districts en amont, elle explore le mécanisme de revenu en testant si les chocs pluviométriques prédisent l'incidence des émeutes dans les districts en aval mais pas dans les districts en amont. Formellement, elle teste ces hypothèses :

- X→Y dans les endroits où X est connu pour affecter M [Les chocs pluviométriques augmenteront les émeutes dans les zones où les chocs pluviométriques affecteront négativement les revenus (en amont du barrage).]
- X n'a pas d'effet sur Y dans les endroits où X n'a pas d'effet sur M [Les chocs pluviométriques n'auront aucun effet sur les émeutes dans les zones où le revenu n'est pas sensible aux précipitations (en aval du barrage).]

Cependant, elle a constaté que la relation entre les chocs pluviométriques et l'incidence des émeutes était tout aussi étroite dans les districts en aval où le revenu n'était pas sensible aux précipitations. Elle interprète ce résultat comme une preuve "suggestive" *contre* le mécanisme de revenu. Pour être clair, Sarsons n'a mené aucune analyse de médiation : elle n'a pas mesuré le revenu de chaque village et quantifié l'effet direct des chocs pluviométriques sur les émeutes et l'effet indirect des chocs pluviométriques sur les émeutes par le biais des revenus. Au lieu de cela, elle a recherché un effet du traitement hétérogène que la théorie aurait impliqué et, n'en trouvant aucune preuve, a conclu que le mécanisme de revenu peut être moins important qu'on ne le pensait auparavant.

Ce que vous pouvez faire... Dans les projets futurs, demandez-vous : si le mécanisme est M, pour quels groupes ou unités s'attendre à un effet de traitement, et pour quels groupes ou unités ne pas s'attendre à une réponse au traitement ? Ensuite, testez si ces prédictions sont étayées par vos données et interprétez cela comme une preuve suggestive pour ou contre le mécanisme proposé M. Gardez à l'esprit que de telles preuves ne sont pas décisives car les groupes pourraient différer d'autres manières qui pourraient affecter leur réaction au traitement.

Nous pouvons également rechercher des preuves suggestives en examinant l'effet de notre traitement sur divers résultats.

Encore une fois, bien qu'il soit difficile de quantifier les preuves d'un mécanisme à l'œuvre, nous pouvons toujours explorer les implications vérifiables de la théorie dans laquelle notre mécanisme apparaît. Au point 7, nous l'avons fait en se demandant si le traitement a affecté des sous-groupes particuliers pour lesquels

[7]Heather Sarsons, "Rainfall and Conflict: A Cautionary Tale." Journal of Development Economics 115 (2015): 62-72.

un effet de traitement est induit par notre théorie. Une autre approche consiste à explorer si le traitement affecte uniquement les résultats induits par notre théorie.

Par exemple, de nombreux chercheurs en sciences sociales s'intéressent à la manière dont l'éducation de masse influence la démocratie. Plusieurs théories de la démocratisation s'attendent à ce que différents mécanismes relient l'éducation et la démocratie. Premièrement, selon la théorie de la modernisation, l'éducation pourrait faciliter le bon fonctionnement de la démocratie en sapant les attachements au groupe (comme l'appartenance ethnique ou la religion) au profit du mérite.[8] Deuxièmement, selon les théoriciens sociaux de l'oppression, l'éducation pourrait saper la démocratie en renforçant l'obéissance à l'autorité, qui est inhérente à une structure de classe.[9] Troisièmement, selon de nombreux politologues et psychologues, l'éducation peut encourager la participation démocratique en donnant aux individus la capacité d'acquérir et d'agir sur la connaissance.[10] Friedman et al. (2011)[11] décident de séparer ces mécanismes en enquêtant sur les résultats d'une expérience de terrain dans laquelle des filles kenyanes ont été assignées de manière aléatoire pour recevoir une subvention à l'éducation. Ils ont effectué un suivi auprès des élèves cinq ans après le programme et leur ont posé plusieurs questions visant à tester lequel de ces trois mécanismes était à l'œuvre : les filles ont-elles accepté le droit d'un mari de battre sa femme ? Un parent a-t-il participé au choix de son mari? Dans quelle mesure la fille s'est-elle identifiée à son groupe religieux ou ethnique ? La fille lisait-elle régulièrement les actualités ?

Le tableau suivant décrit la direction des effets que chaque théorie suggérerait. Notez que les divers mécanismes testés ici résultent de théories avec des prédictions divergentes sur certains de ces résultats. Les prédictions de chacun des trois mécanismes sont décrites sur les lignes, suivies des résultats réels. Nous pouvons voir que deux des résultats recueillis appuient la théorie de la modernisation. Cependant, la théorie de la modernisation aurait prédit une diminution de l'appartenance aux groupes religieux ou ethniques (en réalité, il n'y avait aucun effet) et n'aurait eu aucune prédiction pour la lecture des actualités (en réalité, la lecture des actualités a augmenté). Aucune des prédictions du mécanisme d'obéissance à l'autorité n'était étayée par les données. Cependant, les données corroboraient les quatre prédictions de la théorie de l'autonomisation individuelle. Les auteurs concluent qu'il est plus probable que $X ß M3 ß Y$ que $X ß M1, M2 ß Y$.[12]

Mécanisme	Acceptation du droit du mari de battre sa femme (Y1)	Parent impliqué dans la sélection du mari (Y2)	Association avec la religion, l'identité ethnique (Y3)	Lit les actualités (Y4)
(M1) Modernisation	↓	↓	↓	Pas d'effet
(M2) Obéissance à l'autorité	↑	↑	↑	Pas d'effet
(M3) Autonomisation individuelle	↓	↓	Pas d'effet	↑
Effet réel	↓	↓	Pas d'effet	↑

[8]Marion Joseph Levy, Modernization and the Structure of Societies, Princeton University Press, 1966.

[9]Frantz Fanon, The Wretched of the Earth, Grove Press, 1964. John Lott, Jr., "Public Schooling, Indoctrination and Totalitarianism," Journal of Political Economy 107(6), 1999.

[10]Gabriel Almond et Sidney Verba, The Civic Culture: Political Attitudes and Democracy in Five Nations, Sage Publications, 1963. Robert Mattes et Michael Bratton, "Learning about Democracy in Africa: Awareness, Performance, and Experience," American Journal of Political Science, 51(1), 2007.

[11]Willa Friedman, Michael Kremer, Edward Miguel et Rebecca Thornton, "Education as Liberation?" NBER Working Paper 16939, 2011.

[12]Dans l'étude proprement dite, les auteurs ont été surpris de découvrir des preuves que l'éducation augmentait également l'acceptation de la violence politique par les individus. S'ils soutiennent toujours que l'autonomisation individuelle est responsable de la relation entre l'éducation et la démocratie, ils avertissent que l'éducation ne conduit pas toujours à la démocratisation (c'est-à-dire M3→Y mais il est également possible que M3→NON Y). Néanmoins, leur approche est une démonstration utile de la façon dont de multiples résultats peuvent éclairer les mécanismes.

Cette étude, comme l'étude de Sarsons, n'essaie pas de quantifier dans quelle mesure l'effet de X sur Y est transmis via M. Cependant, grâce à une enquête approfondie sur divers résultats, les auteurs sont en mesure de fournir des preuves suggestives des mécanismes qui semblent les plus plausibles.

Ce que vous pouvez faire... Dans de futurs projets, posez-vous la question : si le mécanisme est M, devrais-je m'attendre à ce le traitement affecte certains résultats ? Devrais-je m'attendre à ce le traitement n'affecte pas certains résultats ? Ensuite, testez si ces prédictions sont étayées par vos données et interprétez cela comme une preuve suggestive pour ou contre le mécanisme proposé M.

L'élaboration de traitements complexes peut aider à comprendre quelle partie du traitement "produit l'effet".

Parfois, les chercheurs expérimentaux essaieront de mieux comprendre les mécanismes en ajoutant ou en soustrayant des éléments du traitement qui sont censés déclencher différents mécanismes. Cette approche est parfois appelée "analyse de médiation implicite" car différents composants de X sont censés manipuler *implicitement* certains mécanismes. Ceci, bien sûr, est une hypothèse : parce que nous ne mesurons pas M directement, nous nous appuyons sur une affirmation théorique selon laquelle le composant A déclenchera M, alors que le composant B ne le fera pas.

Par exemple, de nombreux gouvernements, dont le Mexique, le Brésil, la Tanzanie ou l'Ouganda, ont créé des programmes de transferts monétaires conditionnels pour lutter contre la pauvreté. Ces programmes fournissent de l'argent aux personnes pauvres, mais ils sont souvent assortis de conditions telles que la fréquentation d'une école ou un programme de formation professionnelle. Jusqu'à récemment, nous savions seulement que ces programmes (X) réussissaient à réduire la pauvreté (Y) et que X provoquait Y soit via le cash reçu, soit via l'assiduité requise à l'école ou aux programmes d'emploi. Pour distinguer ces mécanismes, Baird et al. (2011)[13] ont mené une expérience au Malawi, où ils ont assigné un groupe de familles pour recevoir un transfert en cash *conditionnel* à la fréquentation scolaire régulière de leurs filles, un autre groupe de familles pour recevoir l'argent *sans condition*, et un groupe de contrôle pour ne recevoir aucun transfert. Ce design manipulait "implicitement" M : alors que les filles du groupe de transfert inconditionnel pouvaient également poursuivre une éducation, la fréquentation scolaire (la condition étudiée) serait probablement plus élevée dans le groupe qui devait la poursuivre. Sans surprise, la fréquentation scolaire et les résultats aux tests étaient meilleurs pour le groupe recevant des transferts monétaires conditionnels. Cependant, leurs mesures de Y - le taux auquel les filles sont tombées enceintes ou se sont mariées - étaient en fait meilleures (inférieures) dans le groupe recevant les transferts monétaires inconditionnels. Les auteurs ont conclu que les exigences de fréquentation associées aux transferts monétaires conditionnels n'étaient probablement pas le mécanisme responsable du succès de ces programmes dans la réduction des symptômes de la pauvreté.

Des études comme celles-ci aident non seulement les chercheurs en sciences sociales à en savoir plus sur les canaux par lesquels X provoque Y, mais aussi les décideurs politiques à explorer et découvrir de nouveaux traitements. Après plusieurs autres études ont confirmé Baird et al. en démontrant les effets remarquables des transferts monétaires inconditionnels, de nombreux gouvernements et organisations ont commencé à mettre en œuvre des programmes de transferts monétaires inconditionnels.

Ce que vous pouvez faire... Dans les projets futurs, posez-vous la question : mon traitement peut-il être "décomposé" en plusieurs bras de traitement, certains qui manipulent implicitement M, et d'autres non ? Envisagez d'utiliser un design factoriel pour identifier les effets des différents bras de traitement. Si vous disposez de suffisamment de puissance statistique, la comparaison des différents bras de traitement vous fournira des preuves suggestives pour ou contre M.

[13]Sarah Baird, Craig McIntosh, Berk Ozler, "Cash or Condition? Evidence from a Cash Transfer Experiment," Quarterly Journal of Economics 126, 2011.

Malgré les difficultés à mesurer empiriquement les mécanismes, il convient de s'y attarder sérieusement mais d'être prudent dans notre langage.

Tenter d'identifier les mécanismes causaux est une entreprise noble. Articuler les mécanismes causaux est ce qui nous permet de mieux comprendre le traitement en "boîte noire" et de comprendre pourquoi et comment certains traitements fonctionnent. Même si les affirmations causales peuvent être (et sont souvent) faites sans preuves d'un mécanisme causal, l'exploration des mécanismes causaux est ce qui nous permet d'étendre la frontière de la recherche et de réévaluer la façon dont nos preuves correspondent à nos théories. Pour ces raisons, le public (qu'il s'agisse du grand public ou des relecteurs universitaires) est souvent naturellement impatient que vous exposiez les mécanismes causaux après avoir démontré des preuves d'une assertion causale provocatrice. En prévision de cela, il convient de se demander s'il est possible de concevoir un moyen de tester les mécanismes causaux avant de mener une expérience. Si ce n'est pas le cas, demandez-vous si certaines mesures de résultats ou un traitement par interaction de covariables apporteraient du crédit à un mécanisme causal particulier, et soyez explicite sur les limites de ce type d'analyse dans votre article. Les mécanismes sont un domaine d'investigation passionnant et doivent être pris en compte à la fois dans le design et l'analyse d'une expérience, mais nous devons nous assurer de discuter des mécanismes avec une prudence adaptée à notre capacité à identifier un mécanisme particulier et à éviter de le surestimer.