Approches Comportementales Pour La Mesure des Préférences Morales

Jean-François Bonnefon

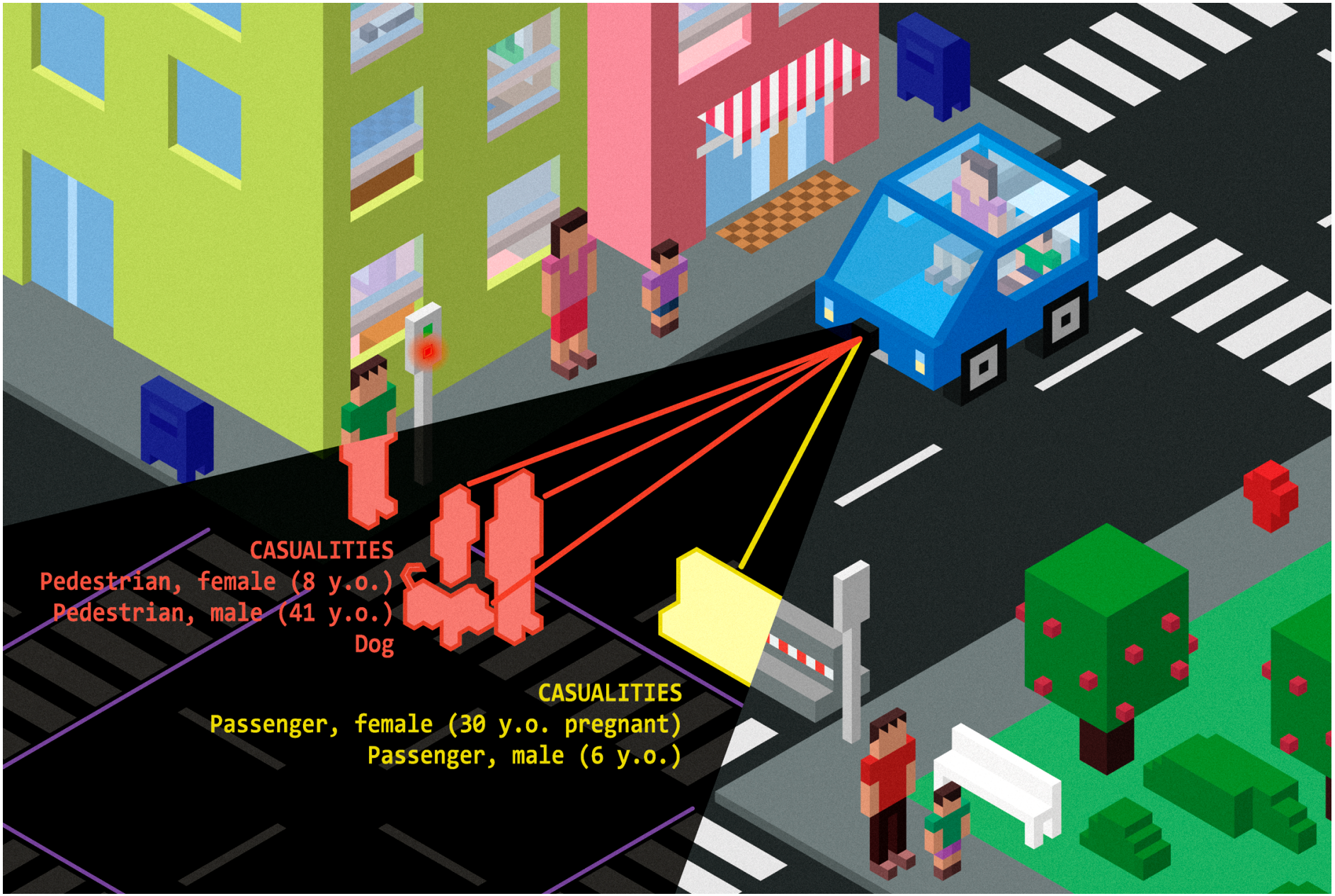
Toulouse School of Economics and Quantitative Social Sciences

Pas de définition consensuelle des préférences « morales »

Philosophie	Préférences impliquées dans la distinction entre les bonnes actions et les actions mauvaises
Économie	Préférences portant sur autre chose que (ou incompatibles avec) les incitations matérielles personnelles
Sociologie	Préférences qui résultent de l'internalisation de normes collectives dont les transgressions font l'objet de réprobation
Psychologie	Préférences dont la non-satisfaction se traduit par des émotions telles que la honte, la culpabilité, l'indignation, ou le dégoût
Biologie	Préférences qui ont été sélectionnées par l'évolution du fait de leur effet positif sur la coopération

En pratique

- Distribution d'une resource critique rare
  Qui prioriser pour les vaccins, les lits de réanimation, la sécurité routière ?
- Valuation d'une situation complexe
  Comment combiner des impacts sur la biodiversité, la vie privée et les inégalités ?
- Approche comportementale descriptive
  Quelles sont (et non quelles devraient-être) les préférences d'un certain groupe ?
- Méthode expérimentale
  Avec des déclinaisons différentes selon les situations


CASUALITIES
Pedestrian, female (8 y.o.)
Pedestrian, male (41 y.o.)
Dog
CASUALITIES
Passenger, female (30 y.o. pregnant)
Passenger, male (6 y.o.)

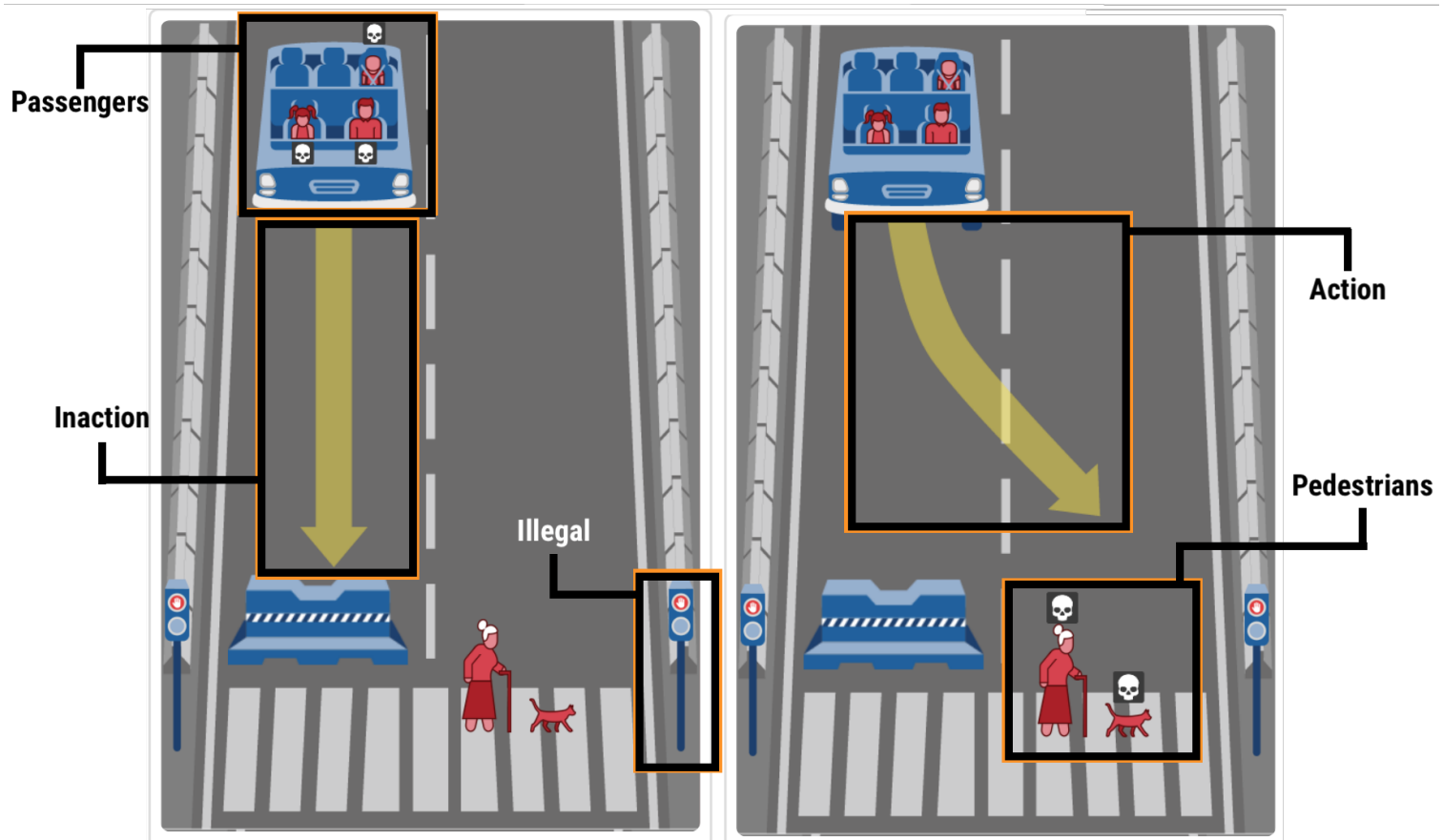
Le besoin de mesurer les préférences

De qui [les voitures autonomes] vont-elles sauver la vie ? Il y a de nombreuses questions éthiques que nous devons résoudre en tant que société. (Ford)

Nous mettrons en œuvre à la fois le cadre légal et ce qui est jugé socialement acceptable. (Daimler AG)

Les décideurs politiques et l'industrie doivent trouver des réponses en collaboration avec le public. (EU)

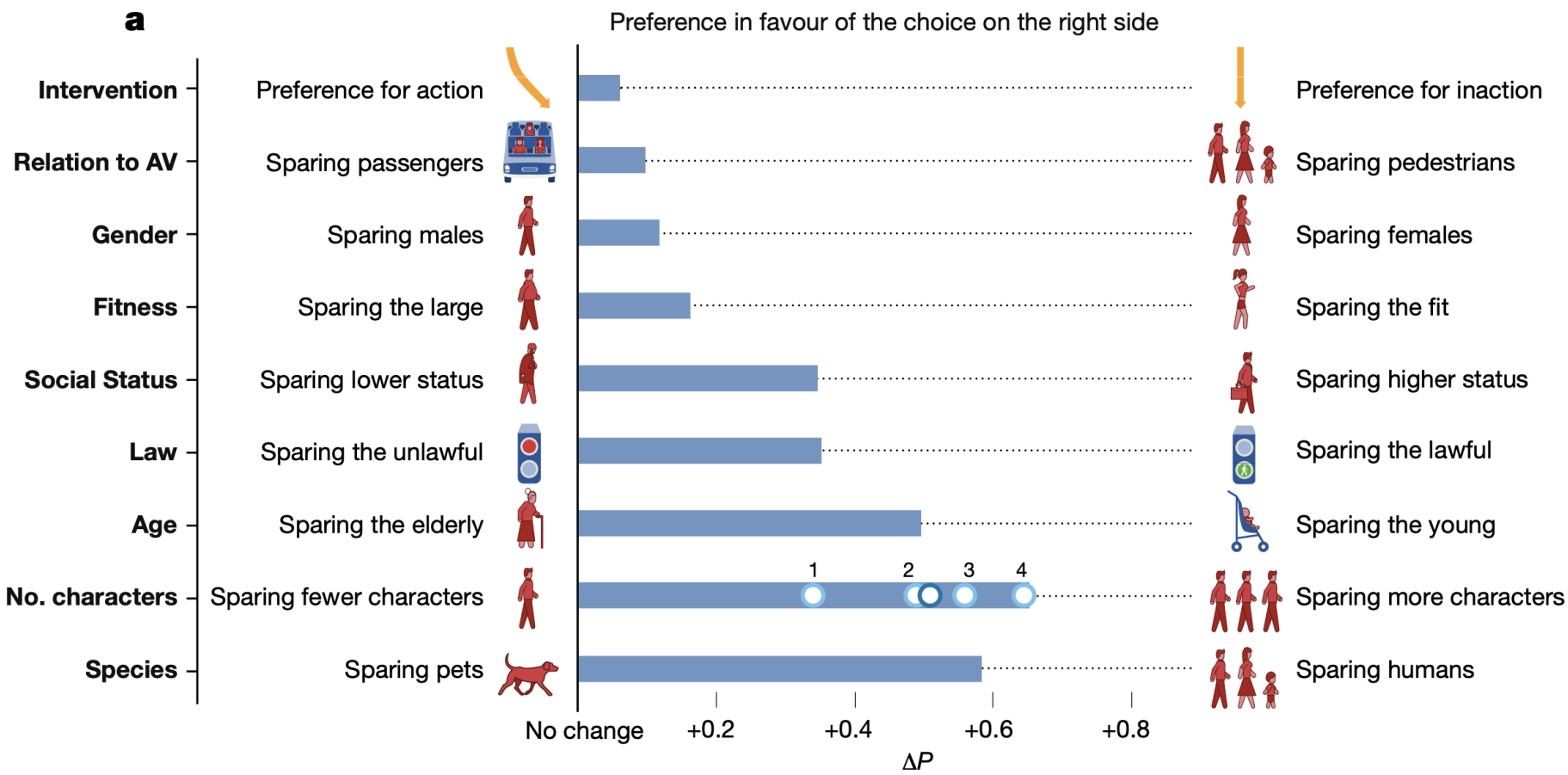
L'expérience Moral Machine



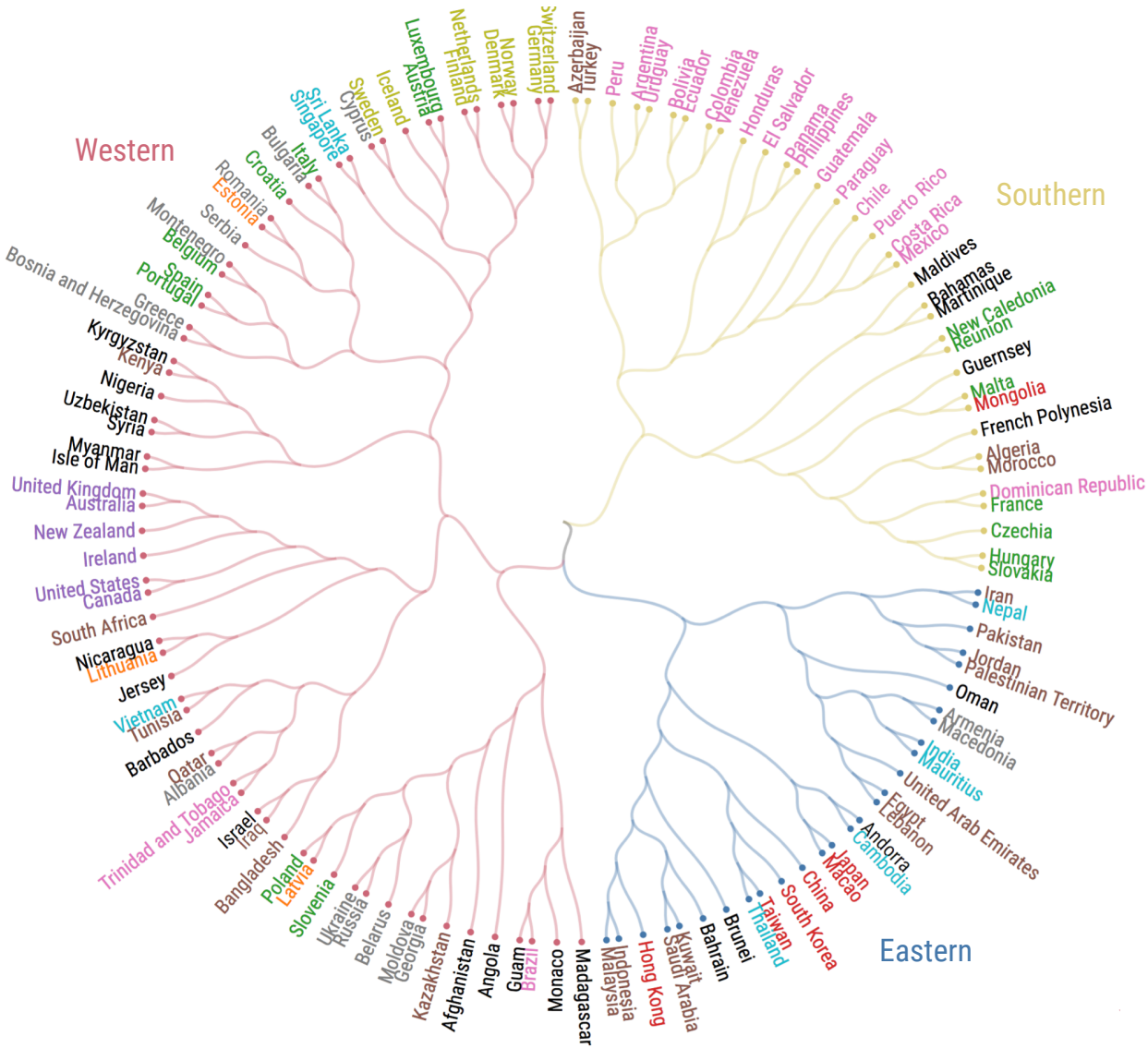
- 10M participants, 100M décisions, 200 pays

  La plus grande étude jamais menée

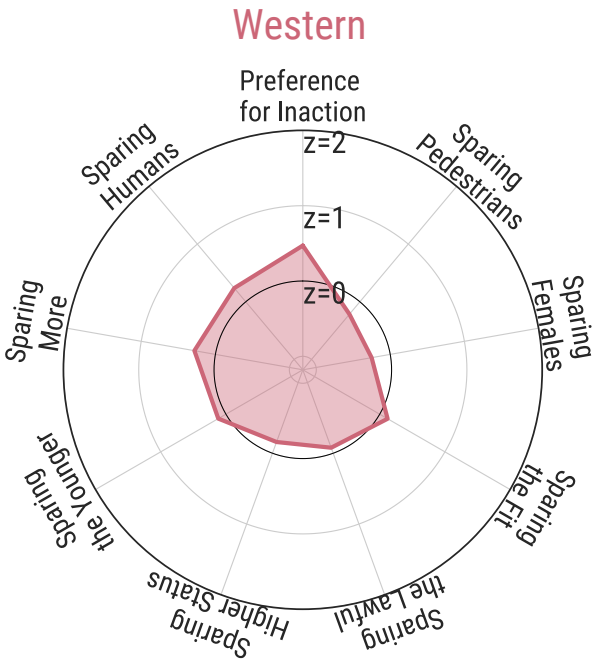
Les préférences globales

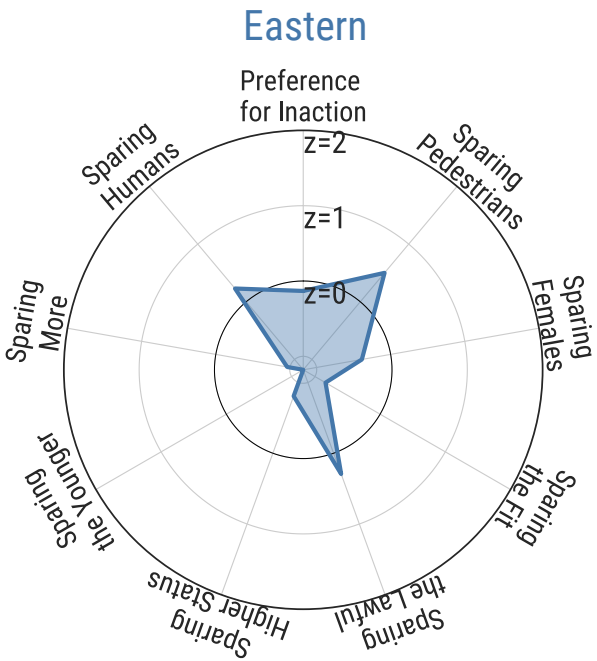


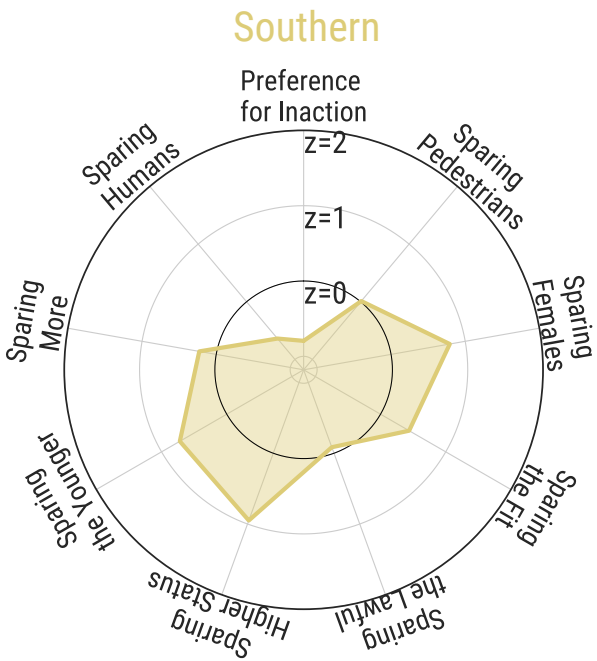
Regroupements culturels (1)



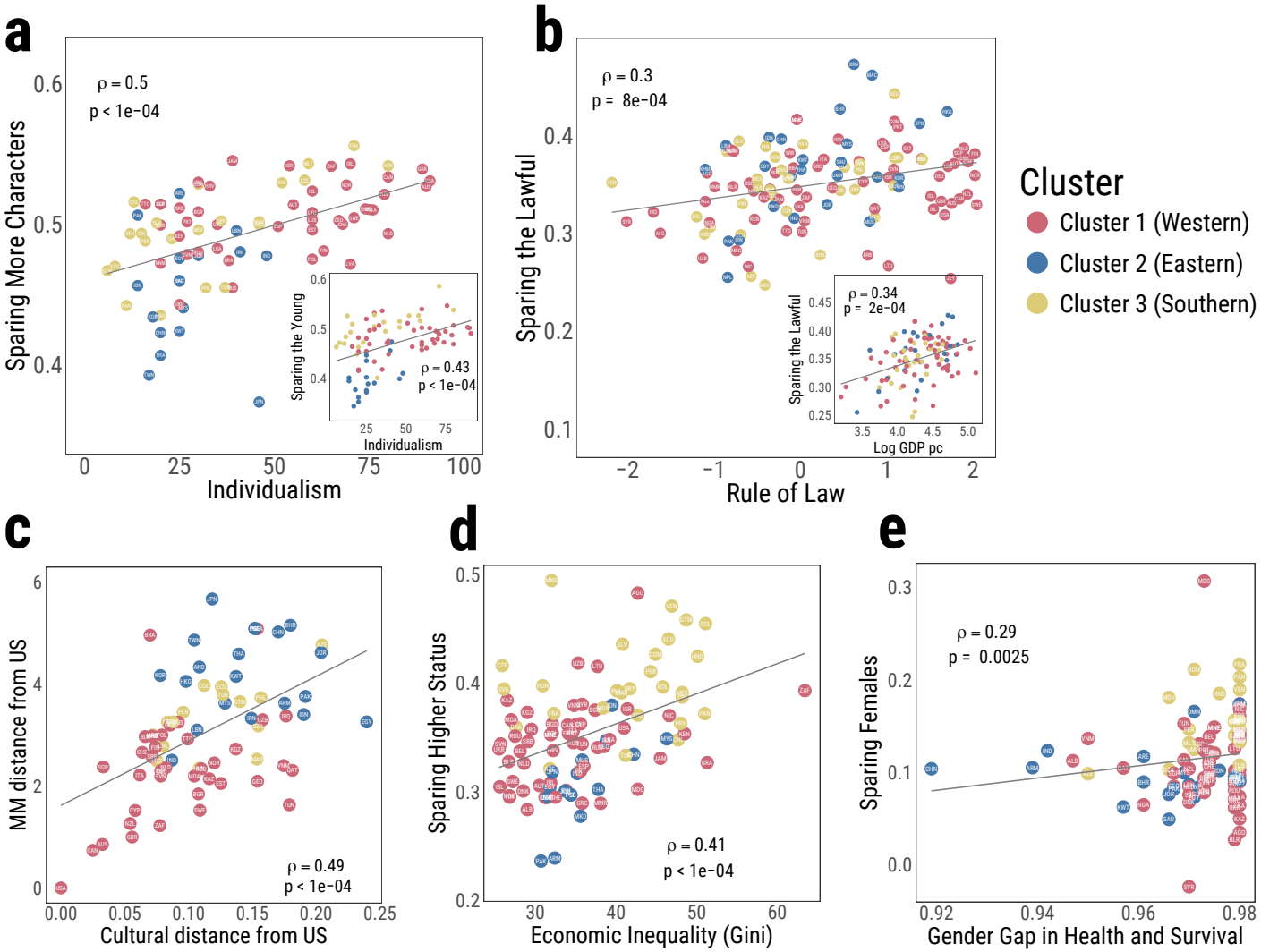
Regroupements culturels (2)



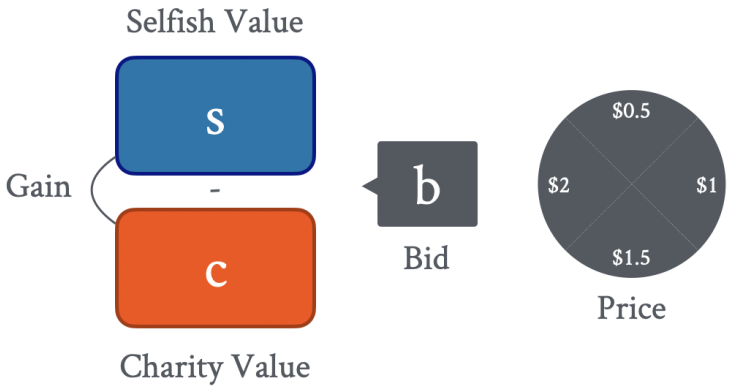




Corrélations culturelles

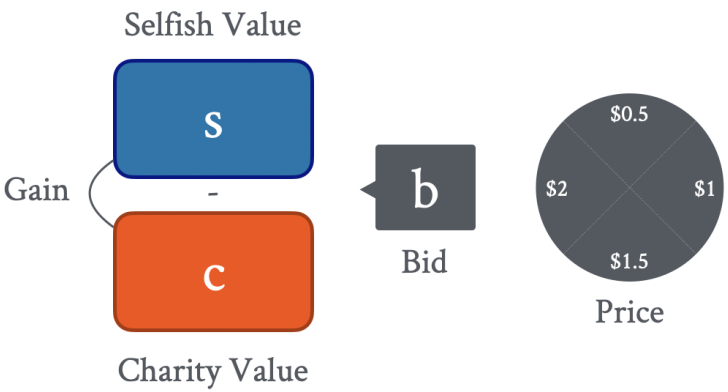


Mesure des préférences morales des investisseurs



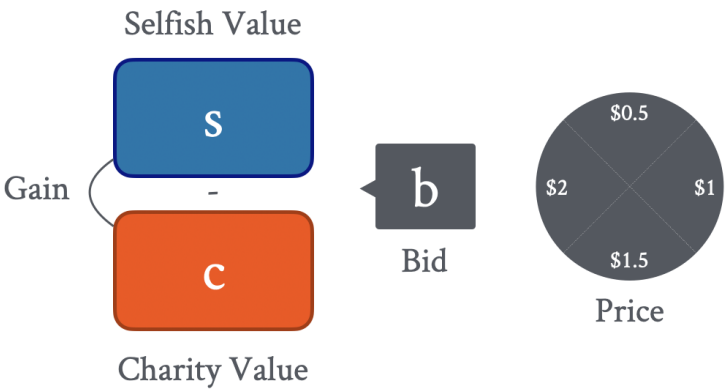
1. L'achat d'un titre peut permettre un profit $s - c$ à l'investisseur
2. c est une valeur distribuée (ou retirée) à une œuvre caritative
3. L'investisseur fait une offre b
4. Un prix p est tiré au hasard. Si $b < p$ pas d'achat, sinon achat au prix p
5. Nous mesurons l'impact de c sur b

Example (1)



1. L'entreprise fait un profit de \$1
2. Si vous n'achetez pas le titre, votre portefeuille ne change pas et l'entreprise verse \$0.3 à l'American Cancer Society
3. Si vous achetez le titre, l'entreprise verse \$0.3 à l'American Cancer Society et met les \$0.7 restants dans votre portefeuille
4. Quelle est votre enchère ?

Example (2)



1. L'entreprise fait un profit de \$1
2. Si vous n'achetez pas le titre, votre portefeuille ne change pas et l'entreprise retire \$0.3 à l'American Cancer Society
3. Si vous achetez le titre, l'entreprise met ces \$1 dans votre portefeuille, puis retire \$0.3 à l'American Cancer Society et les ajoute à votre portefeuille
4. Quelle est votre enchère ?

Pivot ou non

1. L'entreprise fait un profit de \$1
2. Si vous n'achetez pas le titre, votre portefeuille ne change pas et le portefeuille de l'American Cancer Society ne change pas
3. Si vous achetez le titre, l'entreprise met ces \$1 dans votre portefeuille, puis retire \$0.3 à l'American Cancer Society et les ajoute à votre portefeuille
4. Quelle est votre enchère ?

1. L'entreprise fait un profit de \$1
2. Si vous n'achetez pas le titre, votre portefeuille ne change pas et l'entreprise retire \$0.3 à l'American Cancer Society
3. Si vous achetez le titre, l'entreprise met ces \$1 dans votre portefeuille, puis retire \$0.3 à l'American Cancer Society et les ajoute à votre portefeuille
4. Quelle est votre enchère ?

Variations

Pivot

Comme décrit précédemment

Choix de l'œuvre

Random ou parmi ACLU, WWF, American Cancer Society, Food for the Poor, Save the Children

Don direct

Option de donner une partie des gains à l'issue de l'expérience

Population

Restreinte ou non aux utilisateurs de plateformes de trading

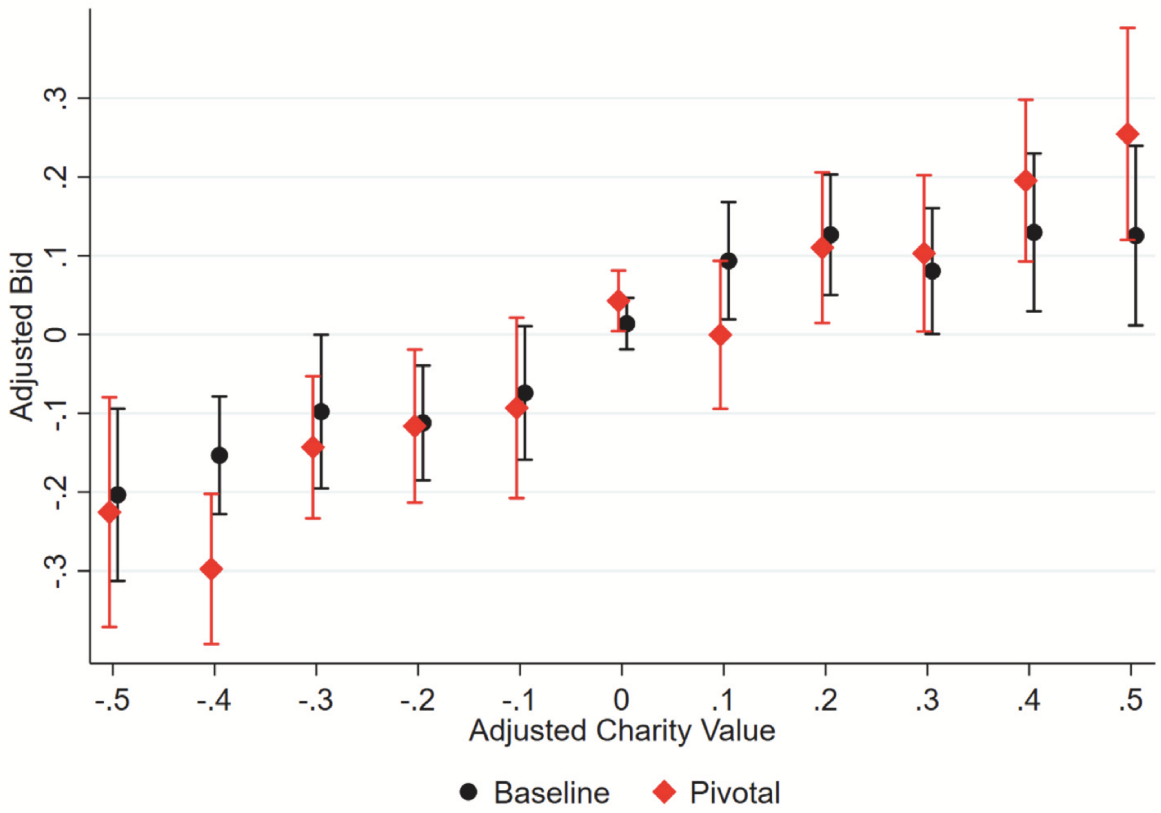
Montants

De l'ordre de \$1 ou \$5

Compréhension

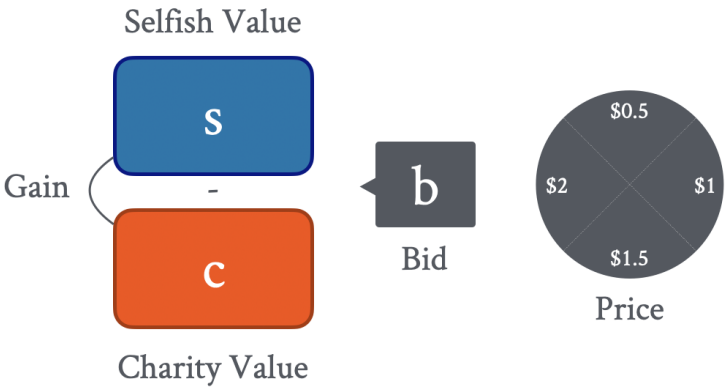
Score parfait au quiz pré-décision

Résultats typiques



- Les participants intègrent 40 à 60% de c dans leur enchère
  Être pivot n'affecte pas cette valuation

Résumé



1. Les enchères b ne reflètent pas seulement s
2. Elles augmentent ou baissent de $\sim$ 50% de c
3. Même en l'absence d'un rôle pivot
4. Les participants payent pour l'alignement
5. Mais pas pour l'activisme

Pour conclure

- Il est parfois nécessaire de mesurer des préférences morales
  Pour construire des instruments, des produits, des politiques
- On mesure ces préférences en présentant des arbitrages
  Et en faisant varier systématiquement les variables d'intérêt
- Les méthodes spécifiques sont adaptées au cas d'usage
  Ex. analyse conjointe, disposition à payer
- Mais elles sont remarquablement flexibles
  Benéfices sociaux, lits de réanimation et vaccins, usages IA générative, etc.
- Et utilisables sur les LLMs comme sur les humains
  cf. machine behaviour

Approches Comportementales Pour La Mesure des Préférences Morales

Jean-François Bonnefon

Toulouse School of Economics and Quantitative Social Sciences