在多元文化的样本中情境因素塑造了东方、南方和西方国家电车困境中的道德判断实验数据分析的复刻

小组 2：时天轲、叶子芸、欧星宇

摘 要 本文对在多元文化的样本中情境因素塑造了东方、南方和西方国家电车困境中的道德判断实验数据的分析进行了复刻，主要复刻了研究一的数据分析。本文为研究一的数据分析提出了一个新的视角，即选择离散型的先验分布而不是连续型的先验分布，尝试进行数据分析，并将数据分析的过程和结果与原文献进行比较，试图找到本文的数据分析方法和原文献的相似点和不同点。

关键词 个人力量；意图；道德两难困境；多元文化

1 研究背景

道德两难是伦理学中表示道德冲突和道德困惑的用语，指人们面对复杂的道德情境和交叉性的道德价值网络往往很难分清主次，无法选择，或者说选择任何一种方案都无法满足自己道德上的需求。它常常是同一道德体系内不同道德原则、道德要求之间冲突的集中反映。道德两难困境（trolley problems）指包含道德两难的问题。

在面对道德两难困境时，人们通常持有功利主义和义务论两种倾向。持有功利主义倾向的个体更加关注在道德两难困境中的利益最大化，例如拯救更多的生命等；而义务论倾向的个体则更加关注在道德两难困境中的行为是否道德和符合人道主义，例如个人权利义务的实现等。

个人力量（personal force）指自发地杀害受害者并拯救更多的人。意图（intention）指行为带来的副作用是有意的还是无意的。Greene 等人认为，受到个人力量和意图交互作用的影响，人们面对不同场景的选择是不同的。例如，在岔路口问题（如图 1）中，电车原本的驶向是在铁轨上的五个人，人们可以选择拉动拉杆使电车撞向另一条线路上的一个人。对于功利主义者来说，为了拯救更多生命，使利益最大化，他会选择拉动拉杆；而对于义务论者来说，拉动拉杆侵犯了那个可怜人的权利，这是一种变向的谋杀，因此他不会选择拉动拉杆。

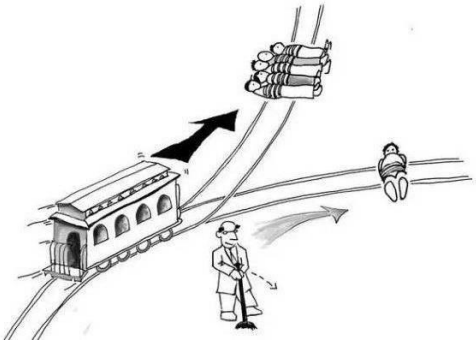


图 1 岔路口问题

而在人行天桥问题（如图 2）中，人们必须把另一个人推下电车前面的人行天桥，这个人会死，但会让电车停下来，挡住电车的 5 个人会得救。相比拉杆问题，人们在人行桥问题上不太会选择功利主义，也就是把人推下去。

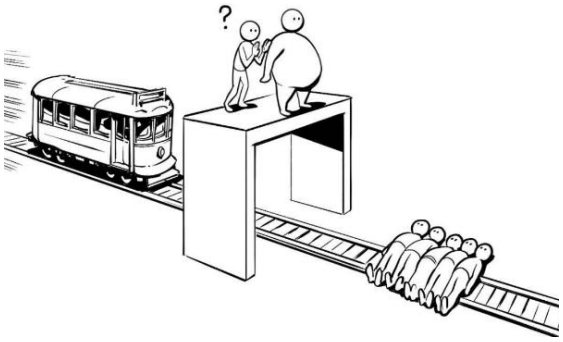


图 2 人行天桥问题

双重效应学说指对于人们来说，无目的的伤害在作为好结果的副作用时是可以接受的。然而，Greene 等人认为，功利主义反应率的差异不能简单地用双重效应学说来解释。他们提出了意图（即伤害作为手段或副作用，指的是双重效应学说）和个人力量（即行为者是否必须使用个人努力杀死受害者并拯救更多人）在道德评级上相互作用的证据。

2 研究问题

原文献的研究检验了三个跨文化假设：1、个人力量对道德判断的影响在文化上是普遍的；2、个人力量和意图对道德判断的交互效应在文化上是普遍的；3、集体主义-个人主义对个人力量和意图影响道德判断的程度具有调节作用，其影响在更集体主义的文化中更强。

在原文献中，实验分为三个部分：第一部分复刻了 Greene 等人的研究 1，

测试了个人力量在道德判断中作用的普遍性；第二部分复刻了 Greene 等人的研究 2，测试了个人力量和意图对道德困境判断交互效应的普遍性；第三部分检验了集体主义缓和了意图和个人力量影响的假设。

3 研究方法

3.1 被试选择

原文献的研究选取了来自 45 个国家的 140 个实验室的 27502 名参与者（17961 名女性，7956 名男性，平均年龄 26.0 岁，标准差 10.3 岁；研究 1：7744 名参与者，4329 名女性，2487 名男性，平均年龄 26.8 岁，标准差 11.1 岁；研究 2：19340 名参与者，13632 名女性，5469 名男性，平均年龄 25.8 岁，标准差 9.98 岁）参与研究。

3.2 测量工具

原文献的研究使用了自编的问卷，主要包含以下几个部分：道德两难困境问题（包含 6 个电车难题和 6 个快艇难题）、牛津功利主义量表、个人主义-集体主义量表、宗教量表、材料理解程度测试（测试被试是否认为材料混乱、问题描述不真实）、粗心测试（包含 3 个荒谬的问题）、注意力测试、对道德两难困境问题的熟悉程度和人口统计学问卷等。

3.3 研究材料

原文献的研究使用的道德两难困境问题包含 6 个电车难题和 6 个快艇难题。6 个电车难题分为实验 1a 两难问题和实验 2a 两难问题两组：实验 1a 两难问题包含人行天桥杆子问题（需要较强个人力量的两难问题）和人行天桥开关问题（需要较弱个人力量的两难问题）；实验 2a 两难问题包含标准岔路口问题、标准人行天桥问题、环状岔路口问题和人行天桥-岔路口-障碍组合问题。6 个快艇难题为与电车难题一一对应的快艇版本（即将情境中轨道上的人变成落入海里的人，将人行天桥变成了快艇）。

3.4 研究设计

原文献的研究一采用了单因素完全随机实验设计。自变量为个人力量（个人力量强、个人力量弱），因变量为被试打出的道德可接受程度。

3.5 研究过程

先给被试呈现一个电车难题，让被试对所描述的行动进行评价（在道德上是否可以接受、道德可接受程度、决策的理由）；接着给被试呈现一个快艇难题（难题展示的顺序是固定的）；最后让被试完成牛津功利主义量表、个人主义-集体主义量表、宗教量表、材料理解程度测试（测试被试是否认为材料混乱、问题描述不真实）、粗心测试（包含 3 个荒谬的问题）、注意力测试、对道德两难困境问题的熟悉程度和人口统计学问卷等。

4 数据分析

4.1 数据筛选

通过对原文献和原文献研究材料的分析，我们总结出原研究中排除的数据有以下几类：注意力测试失败的数据（即在注意力测试中选择了对材料进行错误释义选项被试的数据）、完全无法理解材料被试的数据、熟悉道德两难问题被试的数据、发生技术故障被试的数据、不使用母语被试的数据、粗心大意被试的数据。通过对以上的数据进行排除，我们得到了最终分析需要使用的数据。通过对比可以发现，我们对预处理数据的筛选结果与原文献完全相同（如图 3）。

```
Final sample
Study_1a: 1569
Study_1b: 1426
Study_2a: 3984
Study_2b: 3513
```

Table 1 Summary of sample sizes and exclusions in all cultural clusters

From: Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample

	Eastern	Southern	Western	All
Reason to exclude				
N without exclusion	3,877	5,333	18,292	27,502
Careless responding	156 (4.0%)	82 (1.5%)	256 (1.4%)	494 (1.8%)
Confusion	752 (19.4%)	658 (12.3%)	1,718 (9.4%)	3,128 (11.4%)
Familiarity with moral dilemmas	1,669 (43.0%)	2,501 (46.9%)	10,332 (56.5%)	14,502 (52.7%)
Technical problem	531 (13.7%)	413 (7.7%)	1,225 (6.7%)	2,169 (7.9%)
Non-native speaker	347 (9.0%)	177 (3.3%)	1,305 (7.1%)	1,829 (6.7%)
Failed attention check (study1a)	720 (18.6%)	943 (17.7%)	1,311 (7.2%)	2,974 (10.8%)
Failed attention check (study 1b)	849 (21.9%)	1,042 (19.5%)	1,336 (7.3%)	3,227 (11.7%)
Failed attention check (study 2a)	1,102 (28.4%)	1,071 (20.1%)	4,900 (26.8%)	7,073 (25.7%)
Failed attention check (study 2b)	1,195 (30.8%)	1,367 (25.6%)	5,528 (30.2%)	8,090 (29.4%)
Final sample				
Study 1a	381	622	566	1,569
Study 1b	327	553	546	1,426
Study 2a	323	690	2,971	3,984
Study 2b	277	576	2,660	3,513

Note. Study 1b and study 2b refer to the speedboat dilemmas. Recall that all of our subjects responded to one trolley and one speedboat dilemma.

图 3

4.2 数据分析的复刻

我们复刻了 Greene 等人的研究 1，测试了个人力量在道德判断中作用的普遍性。本部分给被试呈现的道德两难困境问题为实验 1a 两难问题和实验 1b 两难问题。在每组两难问题中，人行天桥杆子问题需要的个人力量较强，而人行天桥开关问题需要的个人力量较弱。因此，通过比较被试对实验 1a 或实验 1b 中两个道德两难困境问题的道德可接受度评分，可以判断个人力量对道德判断的影响是否具有显著性。

提取东部地区的数据对电车难题（实验 1a 两难问题）进行分析。

假设被试对两个两难问题的道德可接受度评分的均值的先验分布均服从均值为 5，标准差为 0.5 的正态分布，且评分的标准差均为 1。依次设定模型、定义先验和似然、进行 MCMC 采样。接着进行模型诊断，结果如图 4 和图 5 所示。

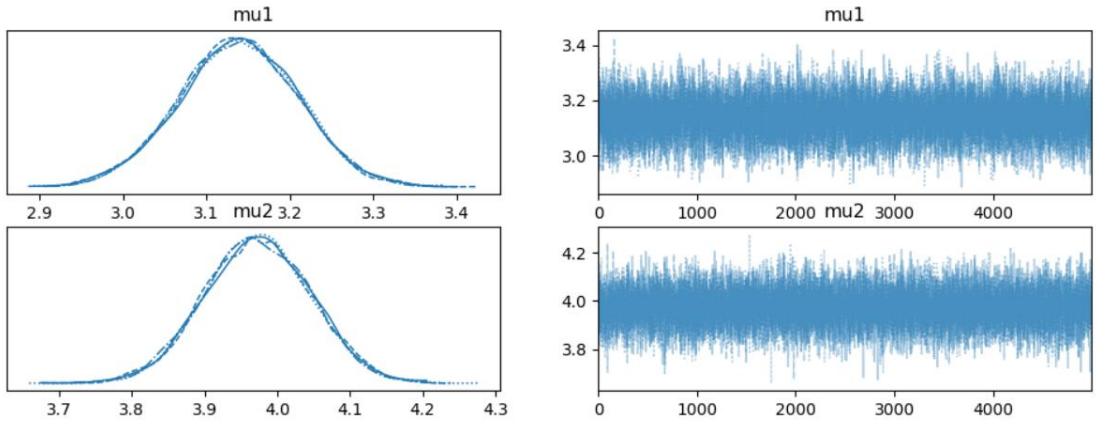


图 4

	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu1	0.000	0.0	20473.0	15299.0	1.0
mu2	0.001	0.0	20191.0	15246.0	1.0

图 5

由图 4 左边的两幅图可得，四条链的后验密度分布保持一致。由图 4 右边的两幅图可得，MCMC 链具有随机性和独立性。由图 5 可得，Rhat 接近 1，表明模型非常稳定。

接着进行先验预测检验，结果如图 6 所示。

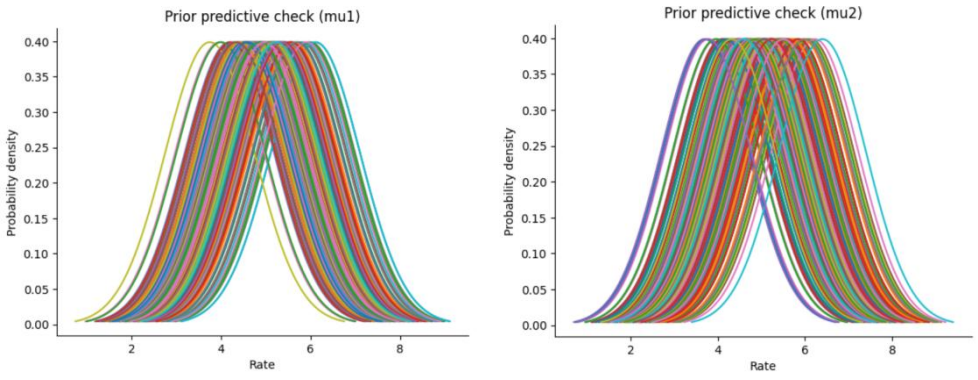


图 6

由图 6 可得，mu1、mu2 的取值集中在道德可接受度评分的范围（1~9 分）内，因此可以认为我们的先验设置比较合理。

接着进行模型评估。依次进行模型拟合、后验预测采样，计算得出预测误差的中位数（MAE）为 1.89，较小。这说明后验模型预测的很准确。

最后进行 t 检验。提出虚无假设 H0：被试对人行天桥杆子问题的道德可接受性评分高于人行天桥开关问题；备择假设 H1：被试对人行天桥杆子问题的道德可接受性评分低于人行天桥开关问题。通过计算，得出被试对人行天桥杆子问题的道德可接受性评分高于人行天桥开关问题的概率接近 0%。因此拒绝虚无假设 H0，接受备择假设 H1，即认为被试对人行天桥杆子问题的道德可接受性评分显著低于人行天桥开关问题。

同时，因为在给被试呈现的材料中，我们要求被试对情境的道德可接受程度进行 1~9 评分，被试只能打出整数的评分，不能打出非整数的评分（例如 3.7），因此分数（rate）是离散型随机变量，应该选择离散型随机变量的分布。正态分布是连续型随机变量的分布，因此我们认为之前先验的设置存在一定的不合理性。

于是我们作出了一些尝试。我们选择多项分布作为先验分布，假设假设被试对两个两难问题的道德可接受度评分对应的概率值服从迪利克雷分布。依次设定模型、定义先验和似然、进行 MCMC 采样。接着进行模型诊断，结果如图 7 和图 8 所示。

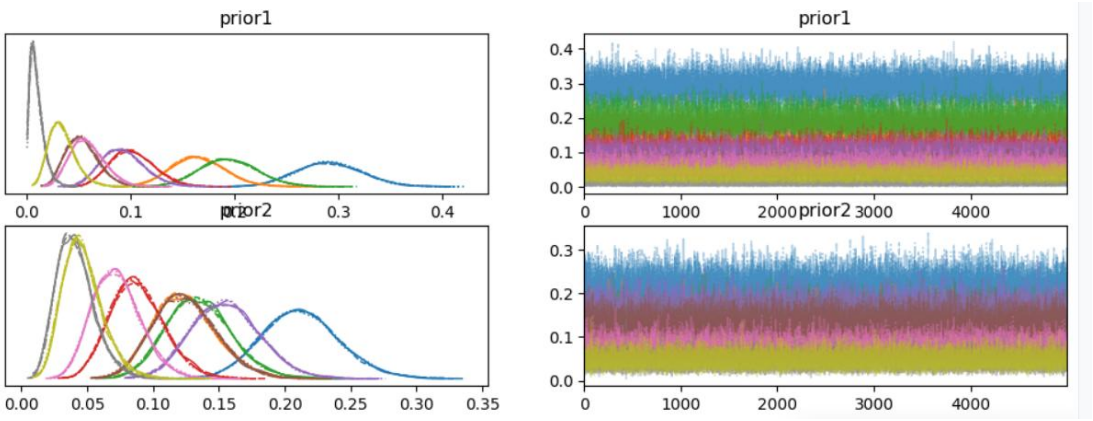


图 7

	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
prior1[0]	0.0	0.0	33435.0	15382.0	1.0
prior1[1]	0.0	0.0	29615.0	13782.0	1.0
prior1[2]	0.0	0.0	29929.0	15167.0	1.0
prior1[3]	0.0	0.0	26079.0	14627.0	1.0
prior1[4]	0.0	0.0	25817.0	14880.0	1.0
prior1[5]	0.0	0.0	22805.0	13998.0	1.0
prior1[6]	0.0	0.0	21652.0	13978.0	1.0
prior1[7]	0.0	0.0	9904.0	8476.0	1.0
prior1[8]	0.0	0.0	18644.0	14354.0	1.0
prior2[0]	0.0	0.0	27186.0	15452.0	1.0
prior2[1]	0.0	0.0	23177.0	14460.0	1.0
prior2[2]	0.0	0.0	22217.0	13319.0	1.0
prior2[3]	0.0	0.0	25323.0	14869.0	1.0
prior2[4]	0.0	0.0	24988.0	16196.0	1.0
prior2[5]	0.0	0.0	23717.0	13340.0	1.0
prior2[6]	0.0	0.0	21874.0	13615.0	1.0
prior2[7]	0.0	0.0	18921.0	13477.0	1.0
prior2[8]	0.0	0.0	25282.0	14334.0	1.0

图 8

由图 7 左边的两幅图可得，四条链的后验密度分布保持一致。由图 7 右边的两幅图可得，MCMC 链具有随机性和独立性。由图 8 可得，Rhat 接近 1，表明模型非常稳定。

接着进行先验预测检验，结果如图 9 所示。

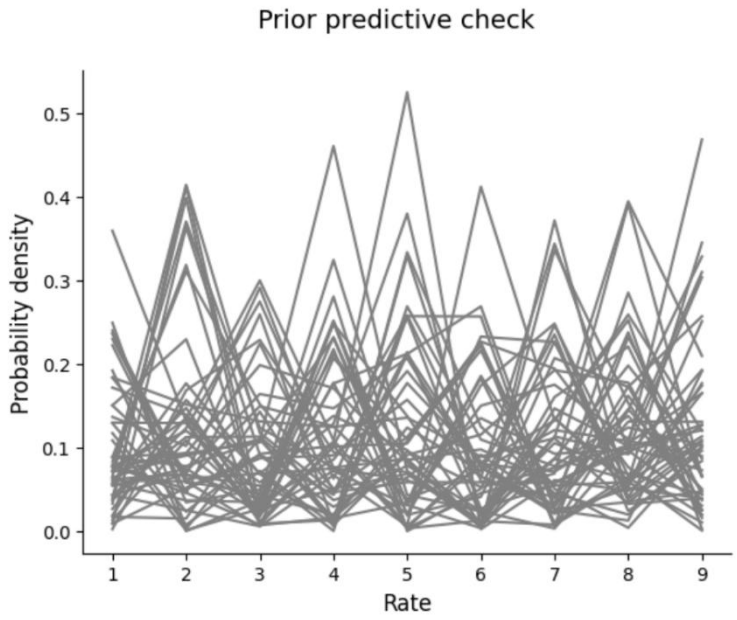


图 9

由图 9 可得，分数（rate）集中在 1~9 之间，且概率密度函数因变量的最大值不超过 1，因此可以认为我们的先验设置比较合理。

接着进行模型评估。依次进行模型拟合、后验预测采样，计算得出预测误差的中位数（MAE）为 1.85 和 1.99，较小。这说明后验模型预测的很准确。接着计算后验预测 HDI，得出所有超过后验预测范围 HDI 的数量均为 0，说明该模型的预测准确率接近 100%。

由于是离散型随机变量，所以不使用 t 检验，我们使用卡方检验。计算得出贝叶斯因子的值为 1.222，较小。这说明被试对人行天桥杆子问题的道德可接受性评分与人行天桥开关问题的差异不显著。

5 结论与反思

通过使用不同的先验分布和不同的检验方式，我们得出了两种截然不同的结果：通过选择正态分布作为先验分布、使用 t 检验作为检验方式，我们得出被试对人行天桥杆子问题的道德可接受性评分显著低于人行天桥开关问题，即个人力量越强，被试的道德可接受度评分越低，这符合原文献的结论和一般常识；而通过选择多项分布作为先验分布、使用卡方检验作为检验方式，我们得出被试对人行天桥杆子问题的道德可接受性评分与人行天桥开关问题的差异不显著，即个人力量对道德判断造成的影响不显著。由此可见，选择不同的先验和统计方式，会

影响对于结论的判定。

在本次作业中我们比较遗憾的是没有使用线性模型、部分池化与完全池化等上课教过的内容，能力没有得到很好的锻炼；我们比较欣喜的是花了很长很长时间对数据进行了筛选，与原文献数据筛选的结果达到了一致，尝试了上课没有教授的使用贝叶斯推论进行 t 检验和卡方检验，并选择了一个离散型随机变量的分布（多项分布）作为先验分布，感受到了不断学习、不断试错和数据本身的魅力。而今迈步从头越，愿我们都能不断学习，无限进步。

参考文献

Bago, B., Kovacs, M., Protzko, J. *et al.* Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nat Hum Behav* **6**, 880–895 (2022). https://doi.org/10.1038/s41562-022-01319-5

Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample

在多元文化的样本中情境因素塑造了东方、南方和西方国家电车困境中的道德判断实验数据分析的复刻

时天轲，叶子芸，欧星宇

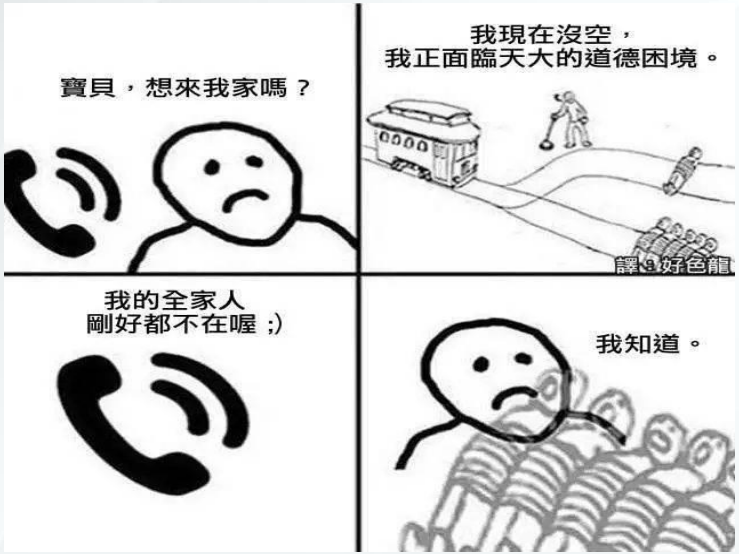
Contents

1

Background & Literature Review

PART ONE

研究背景



道德两难困境 (trolley problems)

指人们面对复杂的道德情境和交叉性的道德价值网络往往很难分清主次，无法选择，或者说选择任何一种方案都无法满足自己道德上的需求。

功利主义倾向

功利主义关注在困境中的利益最大化，比如拯救更多的生命

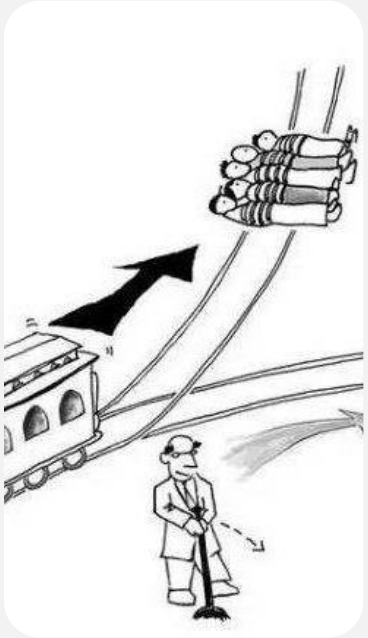
义务论倾向

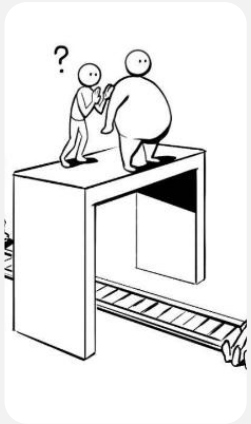
关注行为是否道德、符合人道主义，比如更在乎个人的权利义务等

道德两难问题 (trolley problems)

电车难题

在这个例子中，火车原本的驶向是五个在铁轨上的人，你可以选择拉动拉杆使火车转向创向另一条线路上的一个人





人行桥场景

另一个版本中，一个人必须把另一个人推下电车前面的人行桥，这个人会死，但会让电车停下来，挡住电车的5个人会得救。

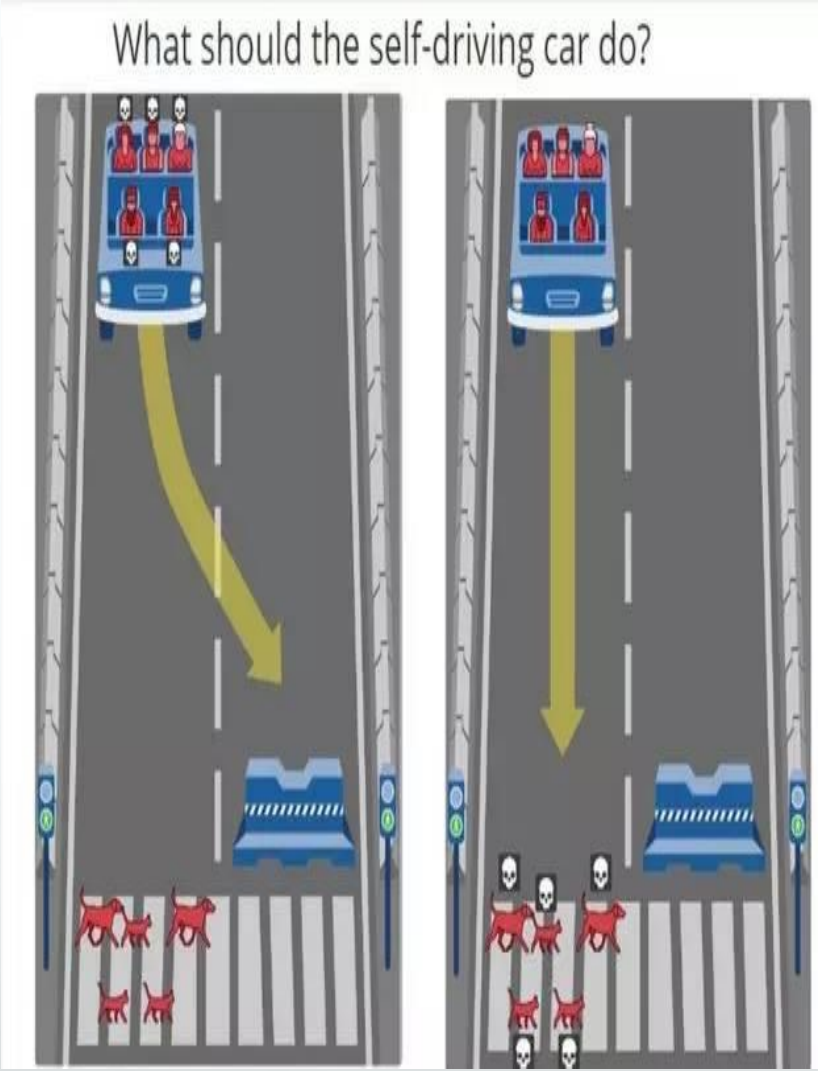
人们对于上述两个场景的选择是不同的，对这个现象，Greene等人认为，这是因为人们收到了个人力量和意图的交互影响。

个人力量（personal force）

个人力量指是否必须使用个人的力量杀死受害者并拯救更多的人，比如把人推下桥

意图（intention）

意图指行为带来的副作用是有意的还是无意的，比如人行桥的副作用则是被推者的死亡，这显然是有意的



研究意义

- ✓ 原始文章具有很大的影响力，但尚未建立可复制性
- ✓ 我们对个人力量和意图对道德判断的文化普遍性的了解是有限的
- ✓ 由此产生的数据库(包含许多类型的电车问题和附加措施)可以帮助和指导未来关于道德思维的研究和应用

2

Study design

PART TWO

研究假设

（1）个人力量对道德判断的影响在文化上是普遍存在的。

（2）个人力量和意图的相互作用对道德判断在文化上是普遍存在的。

（3）集体主义-个人主义对个人力量和意图对道德判断的影响具有调节作用，使得在更具集体主义的文化中这种影响更为强烈。

研究工具

道德困境材料

- ✓ 电车难题（足桥开关、标准足桥、足桥杆、环形、障碍碰撞和标准开关）
- ✓ 快艇问题（足桥开关、足桥杆、环形和标准开关）

文化倾向

- ✓ 个人主义-集体主义四维度量表
- ✓ 牛津实用主义量表（潜在）

道德判断（9点评分）

Is it morally acceptable for Joe to use the pole to push the workman off of the footbridge in order to avoid the deaths of the five workmen, causing the death of the single workman instead?

Please circle one answer: YES NO

To what extent is this action morally acceptable?

Please circle one number:

(Completely unacceptable) 1 2 3 4 5 6 7 8 9 (Completely acceptable)

Please briefly explain why you think this action is morally acceptable/unacceptable:

_____

研究方法

1 道德两难问题

呈现电车难题和快艇难题，难题展示顺序固定

2 注意力测试

- ✓ 关于铁路车厢情境的细节问题
- ✓ 他们是否认为材料混乱
- ✓ 他们是否认为问题的描述真实

3 个人—集体主义问题

- ✓ 牛津实用主义量表
- ✓ 个人主义-集体主义量表
- ✓ 宗教测量

4 人口统计学调查

收入、居住地点、原籍国、移民背景、教育水平、年龄和性别

5 粗心问题

非常基本的问题(例如，“我出生于2月30日”)，错误的回答表明回答粗心

6 熟悉度&技术困难

- ✓ “在进行这个实验之前，您是否熟悉这种类型的道德困境，即通过牺牲一个人的生命来拯救更多人？”
- ✓ 是否有进一步的评论或技术问题

3

Data Analysis

PART THREE

人口学信息

- ✓ 来自45个国家的140个实验室的27,502名参与者（17,961名女性，7,956名男性，平均年龄26.0岁，标准差10.3岁；研究1：7,744名参与者，4,329名女性，2,487名男性，平均年龄26.8岁，标准差11.1岁；研究2：19,340名参与者，13,632名女性，5,469名男性，平均年龄25.8岁，标准差9.98岁）参与研究

数据清洗

- ✓ 按照Greene等人的程序，排除数据：报告说他们发现材料令人困惑的参与者的数据、报告在实验中遇到技术问题的参与者的数据、虚假问题给出错误回答的参与者的数据删除、相应未通过注意力检查问题的数据

分析方法

检验结果是否对贝叶斯分析中选择的先验敏感，指示研究者推断将保持不变的先验范围。该区域的宽度显示了我们的推断对于先验选择的稳健性。

稳健性分析
检验先验

研究1：
配对样本T检验

研究2：
线性回归分析

研究3：
线性混合模型

BF作为指标：强贝叶斯证据的阈值：BF10的决策阈值设为>10用于H1，<1/10用于H0。

代码展示

https://www.heywhale.com/mw-org/NNUPsy/project/658a8264f6e74803f9f24b9e

4

Discussion & Meaning

PART FOUR

研究局限

样本问题

- ✓ 没有针对小规模的狩猎采集社会进行研究
- ✓ 主要由年轻年龄组的受过教育且有互联网接入的个体组成

数据处理

- ✓ 数据收集是在2019年冠状病毒病流行之前和期间进行的，可能以某种方式影响了参与者的回答行为
- ✓ 在主要的确认性分析中排除了81%的样本，这可能导致意外的选择偏差
- ✓ 工作记忆能力较差或文本理解能力较差的人可能更有可能因严格的注意力检查而被排除在外

测量工具

- ✓ 使用单一连续度量的德行-功利主义倾向，这种方法被指责过于简化，无法捕捉更复杂的反应模式。

研究结论

- ✓ 通过使用不同的先验分布和不同的检验方式，我们得出了两种截然不同的结果：通过选择正态分布作为先验分布、使用t检验作为检验方式，我们得出被试对人行天桥杆子问题的道德可接受性评分显著低于人行天桥开关问题，即个人力量越强，被试的道德可接受度评分越低，这符合原文献的结论和一般常识；而通过选择多项分布作为先验分布、使用卡方检验作为检验方式，我们得出被试对人行天桥杆子问题的道德可接受性评分与人行天桥开关问题的差异不显著，即个人力量对道德判断造成的影响不显著。由此可见，选择不同的先验和统计方式，会影响对于结论的判定。

- ✓ 在本次作业中我们比较遗憾的是没有使用线性模型、部分池化与完全池化等上课教过的内容，能力没有得到很好的锻炼；我们比较欣喜的是花了很长很长时间对数据进行了筛选，与原文献数据筛选的结果达到了一致，尝试了上课没有教授的使用贝叶斯推论进行t检验和卡方检验，并选择了一个离散型随机变量的分布（多项分布）作为先验分布，感受到了不断学习、不断试错和数据本身的魅力。而今迈步从头越，愿我们都能不断学习，无限进步。

Thank You

感谢大家观看