

前景理论视角下的诺布效应

Knobe Effect from the Perspective of Prospect Theory

霍雨佳 /HUO Yujia

(山西大学教育科学学院, 山西太原, 030006)
(School of Educational Science, Shanxi University, Taiyuan, Shanxi, 030006)

摘要:对诺布效应进行理论解释一直是实验哲学研究中的一个热点问题。传统的理论认为道德判断对意图判断产生了重要影响,但是这种认识受到了越来越多实验证据的质疑和挑战,特别是在没有道德因素的情景中也发现了诺布效应的存在,这显然是不能用已有的理论来解释的。从前景理论的视角看,对行为结果的得失分析和损失厌恶对意图判断产生了重要影响,这种理论解释不仅与已有的实验结果相符合,而且为重新审视大众直觉的可靠性提供了启示。

关键词:前景理论 诺布效应 意图判断 道德判断

Abstract: The theoretical explanation of Knobe effect has always been a hot issue in experimental philosophy. Traditional theory holds that moral judgment has an important influence on intentional judgment, but this understanding has been questioned and challenged by more and more experimental evidence. Especially in the absence of moral factors, the existence of Knobe effect is also found. Obviously, this can not be explained by previous theories. From the perspective of prospect theory, gain-loss analysis and loss aversion of behavioral outcomes have an important impact on intention judgment. This theoretical explanation not only accords with the experimental results, but also provides enlightenment for re-examining the reliability of popular intuition.

Key Words: Prospect theory; Knobe effect; Judgment of intentionality; Moral judgment

中图分类号: N0 **文献标识码:** A **DOI:** 10.15994/j.1000-0763.2020-07-005

实验哲学给哲学研究带来了革命性的变革,不仅使哲学家可以直接检验自己提出的理论,同时也发现了很多以往坐在“扶手椅”上思辨无法获得的实验结果。诺布效应就是实验哲学带给我们的众多发现之一,也是实验哲学研究的一个典型范例。对诺布效应进行理论阐述更是实验哲学研究中的一个热点问题,本文从前景理论的视角,对其做出一个新的理论阐述,并分析了其与已有理论解释的区别,以及新理论带给我们的启示。

一、诺布效应及现有理论解释的不足

诺布效应(Knobe effect)是在行为意图判断

中表现出的一种不对称现象,由美国实验哲学家约书亚·诺布通过以下实验最早发现并阐释,他给被试呈现以下两个相似的情境:

场景一:破坏环境情境

一个公司的副总裁向董事长汇报说:“现在有一个新方案,这个方案会给公司带来巨额收入,但是也会导致对环境的破坏。”董事长回答说:“我知道这个方案会破坏环境,但是我一点儿都不在乎。我关心的只有我们能赚多少钱,尽快落实吧。”公司最终实施了这个方案,不出所料,这最终导致了环境的破坏。

读完这则实验材料后,被试需要回答的问题是:“这位董事长是故意破坏环境的吗?”

基金项目:全国教育科学规划教育部青年课题项目“不同数感水平儿童的认知差异及干预研究”(项目编号:EBA160405)。

收稿时间:2019年3月19日

作者简介:霍雨佳(1984-)男,山西太原人,山西大学教育科学学院讲师,研究方向为实验哲学、心灵哲学。Email:hyj106@163.com

场景二:改善环境情境

一个公司的副总裁向董事长汇报说:“现在有一个新方案,这个方案会给公司带来巨额收入,而且这个方案的实施还会改善环境。”董事长回答说:“我知道这个方案可以改善环境,但是我一点儿都不在乎。我关心的只有我们能赚多少钱,尽快落实吧。”公司最终实施了这个方案,不出所料,环境因此得到了改善。

读完这则实验材料后,被试需要回答的问题是:“这位董事长是有意改善环境的吗?”

在这两个情境中,董事长的主要目的都是获取利益,方案实施对环境造成的影响只是一种副作用效应。实验结果表明,在损害环境情境下,大部分被试(82%)都认为董事长是有意破坏环境的,但在改善环境情境下,大部分被试(77%)却认为董事长并不是有意要改善环境,^[1]但是,在这两个情境中,董事长在方案实施前都清楚地知道后果,而且对实施方案导致的后果都持不在乎的态度,无论后果是破坏环境还是改善环境,其实都不是他试图实现的,那么,被试为什么会做出两种截然不同的判断呢?

诺布最先对自己发现的这种效应做出了解释,^[2]他认为对行为意图进行判断时包括两个阶段,第一阶段个体会对行为的道德属性进行判断,即判断这个行为是否应该受到惩罚或表扬。第二阶段才对行为意图进行判断,但是在进行意图判断的过程中会参考第一阶段道德判断所得出的结果。当个体已经判断其行为是应该受到惩罚的情况下,个体就会认为其行为是有意为之的。当个体判断其行为既不应该受到惩罚也不应该受到赞扬时,个体就会直接判断其行为是无意为之的。诺布同时还指出第一阶段所进行的道德判断是一种直觉判断,即道德直觉,第二阶段对意图进行的判断则属于更为理性的判断,道德直觉会对理性判断产生重要影响,这种解释一般被称为道德效用理论,强调了道德判断对意图判断的影响。

但是诺布的解释存在以下三个方面的问题,首先,道德效用理论的解释力有限,这个理论只能解释惩罚行为的意图判断和中性行为的意图判断,但是并不能解释赞扬行为的意图判断,比如下面这个情境中的意图判断问题。

Lily和Nancy是好朋友,并一起报名参加了作文比赛,Lily决定帮助Nancy修改她的参赛论文,

她们的另一个朋友Ellen对Lily说:“你帮助Nancy修改她的参赛论文,这样就会降低你自己赢得比赛的机会!”Lily回答说:“我知道帮助Nancy修改她的参赛论文会降低我赢得比赛的机会,但是我一点儿都不在乎。我只想帮助我的朋友。”Lily帮助Nancy修改了她的论文,不出所料,由于Lily的帮助Nancy最终赢得了比赛。

读完这则实验材料后,被试需要回答的问题是:“Lily是有意降低自己赢得比赛的机会吗?”

纳德尔霍夫的实验结果显示有55%的被试认为Lily是有意降低自己赢得比赛的机会。^[3]按照道德效用理论,被试首先会对Lily的助人行为进行道德判断,很显然Lily的助人行为是应该受到赞扬的。但道德效用理论对赞扬行为如何进行意图判断并没有做出解释,所以当个体已经判断其行为是应该受到赞扬的情况下,个体会如何对行为意图进行判断,这仍然是一个有待解决的问题。

其次,诺布的解释指出当个体判断其行为既不应该受到惩罚也不应该受到赞扬时,个体就会直接判断其行为是无意为之的。但是,下面这个“销售业绩”实验的结果却对这种解释提出了挑战,这个实验使用的情景如下:

一个公司的副总裁向董事长汇报说:“现在有一个新方案,这个方案会增加南部地区的销售业绩,但是也会降低北部地区的销售业绩。”董事长回答说:“根据我的计算,北部地区的损失要远远小于南部地区的收益,所以尽快落实吧。”公司最终实施了这个方案,不出所料,这最终增加了南部地区的销售业绩,同时也降低了北部地区的销售业绩。

读完这则实验材料后,被试需要回答的问题是:“这位董事长是有意降低北部地区的销售业绩的吗?”

实验结果显示有80%的被试认为董事长的行为既不应该受到惩罚也不应该受到赞扬,那么按照道德效用理论,大部分的被试就应该判断董事长并不是有意降低北部地区销售业绩的,但是实验结果却表明有75%的被试认为董事长是有意降低北部地区的销售业绩,^[4]这就与诺布的道德效用理论出现了矛盾,表明该理论并不能完全解释诺布效应。

最后,道德效用理论的前提假设就是情境中包含了一定的道德属性,所以被试在做意图判断之前会对行为进行一个道德判断,道德判断的结果进而会影响之后意图判断的结果。但是,有研究表明在

没有道德属性的情境中也存在显著的诺布效应，比如下面这个实验：

场景一：多付的美元

Jason感到自己非常口渴，所以就去自己常去的冷饮店准备买一个大杯的冷饮，在结账前，收银员对他说：“如果今天买大杯的冷饮，要比以前多付1美元”，Jason回答说：“我知道会比以往多付1美元，但是我一点儿都不在乎。我只想要一个大杯的冷饮，赶快结账吧。”服务员给Jason结了账，他得到了一个大杯的冷饮，同时也比以往多付了1美元。

读完这则实验材料后，被试需要回答的问题是：“Jason是有意要多付1美元的吗？”

场景二：免费的杯子

Jason感到非常口渴，所以就去自己常去的冷饮店准备买一个大杯的冷饮，在结账前，收营员对他说：“如果今天买大杯的冷饮，可以得到一个特制的纪念版杯子，用来装冷饮”。Jason回答说：“我知道会得到一个纪念版杯子，但是我一点儿都不在乎。我只想要一个大杯的冷饮，赶快结账吧。”服务员给Jason结了账，他得到了一个大杯的冷饮，同时也得到了一个特制的纪念版杯子。

读完这则实验材料后，被试需要回答的问题是：“Jason是有意要得到一个纪念版杯子的吗？”

实验结果显示有95%的被试认为Jason是有意要多付1美元，而只有45%的被试认为Jason是有意要得到一个纪念杯，^[5]这表明在没有道德背景的情境中诺布效应也是存在的，而道德效用理论却无法对非道德情境中的诺布效应做出解释。

综上所述，道德效用理论并不能对诺布效应做出全面和有效的解释，所以需要引入新的视角来对此效应的成因进行更加深入的探讨。

二、前景理论概述

前景理论是由诺贝尔奖获得者行为科学家卡尼曼及其搭档特沃斯基提出的一种决策理论，被广泛应用在政治、法律、商业和医学等领域的判断和决策中，那么人到底是如何做出判断和决策的呢？为了回答这个问题，他们设计了以下两个情境：

情境1：请在下列两个选项中选择其一：

方案A：你肯定会得到1000元。

方案B：抛一枚硬币，正面朝上，你能得到

2000元，反面朝上，你什么也得不到。

情境2：请在下列两个选项中选择其一：

方案A：确定损失5000元。

方案B：抛一枚硬币，正面朝上，你将没有任何损失，反面朝上，你将损失10000元。

在情境1中，实验结果显示大部分人都会选择方案A，那这是为什么呢？以往认为一般人都是厌恶风险的，所以会选择肯定得到1000元的选项。如果真是这样的话，那么大部分人在情境2中也会选择方案A，因为方案A同样也是无风险的选项。但是实验结果却显示在情境2中大部分人都会选择方案B，而方案B是一个有风险的选项。可见，人们并不是依据风险偏好来进行判断和决策的，所以卡尼曼等人提出了前景理论用以解释上述现象。前景理论认为个体在进行判断和决策之前，会对行为结果的得失进行评估，也就是将结果区分为损失（lose）和获得（gain）两种情况，然后依据结果的得失再做出相应的判断和决策。^[6]在情境1中，得到1000元和得到2000元都是获得，而肯定得到1000元显然更加具有吸引力，所以大部分人都会选择方案A；在情境2中，损失5000元和损失10000元都是损失，因为人们都厌恶损失，所以大部分人会选择放手一搏以避免损失。

那么如何定义事情的得与失呢？前景理论提出事情的得失是和参考点有关系的，对于相同的结果，由于参考点的不同，不同的人可能会得出不同的判断，比如两个学生期末都考了70分，一个同学可能会非常高兴，认为这是一种收获，因为自己上回没及格，所以这回他的目标是及格就可以了，但是结果却比自己的目标还要高。另一个同学可能会十分失望，认为这是一种损失，因为自己成绩一直不错，这回的目标是要考到90分以上，但是结果却比自己的目标低了很多。可以看出，在这个例子中，参考点就是两人的目标，参考点不同，对损失和获得的定义就会不同。此外，现状也是一种常见的参考点，当比现状有所改善和提高时，我们就认为这是获得，比现状还差时，就是损失。最后，参考点也可以是个体的期望。比如你今年的工作业绩不错，经理承诺年底给你一个额外的奖励，但是到了年底经理却没有履行他的诺言，这时你就会感到失望，虽然从表面上看你的现状并没有变坏，但是因为你的期望未能实现，所以你还是感到这是一种损失。

损失和获得给人的感受是不同的，但是两种感

受在程度上有差异吗?为了回答这个问题,他们设计了以下实验:

现在,有一个抛硬币的游戏:

如果是正面朝上,你将赢得150美元,

如果是反面朝上,你将输掉100美元。

这个游戏吸引力大吗?你会参加吗?

根据伯努利的期望效用理论,这个游戏的期望值是25美元,所以你参加这个游戏对你是有利的。但是实验结果却显示大部分人并不愿意参加这个游戏,^[7]这是为什么呢?因为大部分的人都觉得输掉100美元的失落感要远远大于赢得150美元的喜悦感,可见损失比获得给人的感受更加强烈,这种对损失和获得感受的不对称性在生活中也很常见,对于一般人而言,丢失200元的痛苦要高于获得200元的喜悦,实验发现人们对坏事的厌恶和痛苦程度一般是对好事喜悦程度的2倍左右,也就是说如果你女朋友正为丢失了200元而生气时,你至少应给她400元才能平息他的怒火。

综上所述,前景理论主要包括两个方面的内容,一方面,人们在判断和决策时,会根据参考点把事情分为损失和获得;另一方面,人们对损失更加敏感,也就是说损失比获得对人的影响更大。前景理论为判断和决策研究带来了革命性的突破,同时也为解释诺布效应提供了新的视角和启示。

三、诺布效应的前景理论解释

因为前景理论在解释个体判断和决策时具有广泛的适用性,且诺布效用中的意图判断也属于判断中的一种类型,因此可以利用前景理论的基本思想来对诺布效应进行解释。从前景理论的视角看,个体对意图的判断可以分为两个阶段,第一阶段个体会依据参考点对行为的副效应进行判断,并将其分为损失和获得两种情况。第二阶段主要对行为意图进行判断,而且在进行意图判断的过程中会受到第一阶段对副效用判断结果的影响,因为人们普遍对损失更加敏感,所以如果副效应与参考点相比是一种损失,那么个体就会认为其行为是有意为之的,比如在破坏环境情境中,方案未实施之前的环境状态就是参考点,方案的实施破坏了环境,就是一种损失,所以大部分都认为这是董事长故意为之的;如果副效应与参考点相比是一种获得,个体就会认为其行为是无意为之的,比如在改善环境情境中,

改善环境是一种获得,但是董事长却对其并不在乎,所以大部分被试都认为这并不是董事长有意为之的。前景理论与之前的道德效用理论相比具有以下三个方面的优势:

首先,前景理论不但能解释惩罚行为的意图判断,还能解释赞扬行为的意图判断,比如在前面提到的“作文比赛”的场景中,Lily因为帮助Nancy修改论文而降低了自己赢得比赛的机会,这对于Lily来说就是一种损失,所以大部分人都会认为这是Lily有意为之的。可见,前景理论比道德效用理论拥有更强的解释力。

其次,道德效用理论在解释“销售业绩”的场景时出现了与实验证据的矛盾,但是前景理论却可以很好的解释这种场景,在此场景中,两个地区以前的销售业绩就是参考点,由于新方案的实施,产生的副效应就是北部地区的业绩比之前出现了下滑,这显然是一种损失,所以大部分人都认为是董事长有意降低了北部地区的销售业绩。

最后,前景理论不仅可以解释包含道德属性的诺布效应,而且也可以解释不包含道德属性的诺布效应,比如在前面购买冷饮的场景中,Jason购买冷饮时比平时多付了1美元,这就是一种损失,所以大部分人都认为Jason是有意要多付1美元的;而购买冷饮比平时可以多得到一个纪念版杯子,这是一种获得,所以大部分人并不会直接认为这是他有意为之的,而是会结合情境中的其它信息,做出一个综合判断,认为其并不是有意要这个纪念品。可见,前景理论比道德效用理论有更广的解释范围。

前景理论对诺布效应的解释并不依赖于道德判断,当然,前景理论并不是第一个指出诺布效应与道德判断无关的理论,马舍雷提出的权衡假设理论认为意图判断与对成本和回报(cost and benefits)的权衡有直接关系。当副效应被认为是获得回报的一种成本时,个体就倾向于认为副效应是有意为之的;当副效应不能被解释为获得回报的成本时,个体就倾向于认为副效应是无意为之的。^[8]比如在破坏环境情境中,损害环境可以看作是获得回报的一种成本,所以大部分被试都认为董事长是有意要破坏环境的。而在改善环境情境中,改善环境不能解释为公司获得回报的一种成本,因此大部分被试都认为董事长并不是有意要改善环境的。权衡假设理论从表面上看和前景理论非常相似,但其实两种理论之间是有本质差异的,其核心就在于成本

(cost) 和损失 (lost) 是两个非常不同的概念, 一方面, 损失是相对于参考点而定义的, 而成本则不是, 另一方面, 实验结果表明成本和损失这两个概念会对判断产生非常不同的影响, 实验情境如下:

损失情境: 某个游戏, 有 10% 的概率会赢得 95 元, 有 90% 的概率会损失 5 元, 你会参加这个游戏吗?

成本情境: 若某彩票有 10% 的概率会赢得 100 元, 有 90% 的概率什么也得不到, 你会花 5 块钱买这张彩票吗?

你对这两个问题的回答是一致的吗? 实验结果表明大多数被试都会接受成本情境中的提议, 而拒绝损失情境中的提议, 可见成本和损失概念对个体判断的影响并不相同。^[9]这主要是因为大多数人都厌恶损失, 所以损失比成本引起了更加强烈的负面情绪和倾向。但是, 你如果仔细思考, 就可以发现上面两个情境在本质上其实是相同的, 你面临的判断都是要决定——是否要接受得到 95 元或损失 5 元的可能。

权衡假设理论认为个体在进行判断和决策时, 把副效应的结果看成了一种成本, 但这似乎缺乏场景中已知信息的支持, 比如在破坏环境情境中, 董事长对破坏环境, 表现的一点儿都不在乎。这似乎很难让人相信他把破坏环境看成了一种成本。可见, 权衡假设理论对诺布效应的解释力也是非常有限的。

传统的哲学理论认为影响行为意图判断的要素主要有以下三个: 第一是知识或信念, 指个体在实施行为之前是否知道自己的行为会导致怎样的结果。第二是期望, 指个体在实施行为之前是否真的希望行为后果的发生。第三是技能, 指个体是否有能力实施该行为并达到其结果。^[10]但是, 诺布效应的发现却表明意图判断可能并非只受到这三个因素的影响, 比如在破坏环境和改善环境两个情境中, 信念、期望和技能三个要素都是相同的, 但是被试对意图判断的结果还是大相径庭。这就表明一定还有其它一些因素在影响人们的意图判断过程。从前景理论看, 人们对行为结果的得失分析和损失厌恶对意图判断也产生了重要影响, 这也得到了最新实验证据的支持, 美国哈佛大学大卫·莱瓦里发表在《科学》上的实验结果表明, 大脑在进行判断和决策时会自动对事物进行比较,^[11]而得失分析的本质就是将行为结果与参考点相比较, 这种比较无需

消耗认知资源, 是一种内隐的认知操作, 因此个体在判断的过程中根本就不会有所察觉。

四、启 示

诺布效应的前景理论解释表明, 大众在意图判断时并不是完全理性的, 这也促使我们开始反思大众直觉的可靠性。首先对直觉可靠性提出质疑和挑战的是卡尼曼和特沃斯基, 他们用系统的实验证明人类大脑所产生的直觉判断存在一些系统性的错误, 他们将之命名为“启发法和偏见”,^[12]对于什么是“启发法”, 不同的行为科学家有不同的理解, 诺贝尔奖获得者行为经济学创始人理查德·塞勒认为启发法就是经验法则的一种华丽表达,^[13]而卡尼曼认为启发法的本质是将一个难以回答的问题替换成一个相对容易回答问题的过程, 但是我们认为启发式的本质是使用少量信息作出快速判断的认知过程。比如在破坏环境的情景中, 大部分个体直接依据破坏环境是一种损失, 就判断董事长的行为是有意的。人类之所以会使用启发法进行判断, 主要是因为大脑的认知能力是有限的, 诺贝尔奖获得者人工智能专家赫尔伯特·西蒙称之为“有限理性”,^[14]脑科学的实验结果为这个观点提供了有力的证据, 脑成像研究表明大脑在进行决策和判断时遵循“最省力法则”,^[15]即大脑会想法设法使用最少的信息和认知操作来达到节约认知资源的目的。目前, 实验哲学领域有一种趋势, 就是通过调查普通大众对哲学问题的直觉判断来检验哲学家提出的观点和主张。但是, 大众直觉在很多情况下并不可靠, 所以这也使此种验证的价值大打折扣。那么, 实验哲学未来应该朝哪个方向发展呢? 实验哲学的本质是用实验方法解决哲学问题, 传统的哲学家是依靠哲学直觉来解决哲学问题的, 但是哲学直觉是否可靠? 哲学直觉是怎样形成的? 哲学直觉受到哪些因素的影响? 这些问题正是实验哲学家感兴趣的话题。现在, 已有越来越多的实验结果对哲学直觉的可靠性提出了质疑和挑战, 并颠覆了许多传统的哲学观点。诺布和尼克尔斯在《实验哲学宣言》中明确提出: “实验哲学的目标是对概念应用的影响因素进行说明, 并揭示其内在的心理过程和机制。在这里, 不是用精确刻画直觉的实际样式来度量进步, 而是用对直觉的解释深度来衡量进步。”^[16]前景理论对诺布效应做出了更有力的解释, 使我们对人们

是怎样使用“故意”这个概念的影响因素和心理机制有了更加深刻的理解,同时也使我们对大众直觉有了新的认识。

[参考文献]

- [1]Knobe, J. 'Intentional Action and Side-effects in Ordinary Language'[J]. *Analysis*, 2003, 63(279): 190-194.
- [2]Knobe, J. 'Intentional Action in Folk Psychology: An Experimental Investigation'[J]. *Philosophical Psychology*, 2003, 16(2): 309-323.
- [3]Nadelhoffer, T. 'On Praise, Side Effects, and Folk Ascriptions of Intentional Action'[J]. *Journal of Theoretical and Philosophical Psychology*, 2004, 24(2): 196-213.
- [4]Knobe, J., Mendlow, G. 'The Good, the Bad and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology'[J]. *Journal of Theoretical and Philosophical Psychology*, 2004, 24(2): 252-258.
- [5]Mallon, R. 'Knobe Versus Machery: Testing the Trade-off Hypothesis'[J]. *Mind and Language*, 2008, 23(2): 247-255.
- [6]Kahneman, D., Tversky, A. 'Prospect Theory: An Analysis of Decision Under Risk'[J]. *Econometrica*, 1979, 47(2): 263-291.
- [7]Kahneman, D. *Thinking Fast and Slow*[M]. London: Allen Lane, 2011, 283-286.
- [8]Machery, E. 'Understanding the Folk Concept of Intentional Action: Philosophical and Experimental Issues'[J]. *Mind & Language*, 2008, 23(2): 165-189.
- [9]Kahneman, D., Tversky, A. 'Choice, Values, and Frames'[J]. *American Psychologist*, 1984, 39(4): 341-350.
- [10]Shepherd, J. 'Action, Attitude, and the Knobe Effect: Another Asymmetry'[J]. *Review of Philosophical Psychology*, 2012, 3(2): 171-185.
- [11]David, L., Daniel, G., Timothy, W., Beau, S., David, A., Thalia, W. 'Prevalence-induced Concept Change in Human Judgment'[J]. *Science*, 2018, 360(6396): 1465-1467.
- [12]Tversky, A., Kahneman, D. 'Judgment Under Uncertainty: Heuristics and Biases'[J]. *Science*, 1974, 185(4157): 1124-1131.
- [13]Thaler, R. H. *Misbehaving: The Making of Behavioral Economics*[M]. New York: W.W. Norton & Company, 2015, 23-28.
- [14]Simon, H. A. *Reason in Human Affairs*[M]. San Francisco: Stanford University Press, 1983, 22-27.
- [15]Kool, W., McGuire, J. T., Rosen, Z. B., Botvinick, M. M. 'Decision Making and the Avoidance of Cognitive Demand'[J]. *Journal of Experimental Psychology*, 2010, 139(4): 665-682.
- [16]Knobe, J., Nichols, S. *Experimental Philosophy*[M]. New York: Oxford University Press, 2008, 3-14.

[责任编辑 王巍 谭笑]