• 专题：实验哲学 •

编者按：

实验哲学是最近二十年来兴起的一种哲学思潮，主张运用心理学的方法研究传统的哲学问题，这种跨学科研究引起了哲学界的热烈讨论，尤其是概念分析哲学的批评和实验哲学的回应成为哲学的一个非常值得关注的话题。实验哲学呼吁哲学学者重新看待哲学和科学的互动关系，在新的时代追求一种跨学科的研究、合作。

本次专题选取了 4 篇实验哲学文章，梅剑华的《重复危机与实验哲学》是从重复危机对心理学的挑战来考察重复危机是否对实验哲学构成挑战。聂敏理的《哲学与实验——实验哲学的兴起及其哲学意义》从哲学传统尤其是古希腊哲学的角度对实验哲学的价值进行了阐发，通过介绍当代情境主义与德性伦理学之间的论战，表明实验哲学对传统的哲学问题和思想产生重要影响。张学义通过引入功能核磁共振（fMRI）等认知神经科学的技术手段，对心灵理论给出了自己的解释，从实验哲学角度试图回答这一哲学难题。王洪光对实验哲学的方法提出了一个批评，这种批评有助于实验哲学家对于使用调查方法具有高度的自觉。

本次四篇文章视角各有不同，对实验哲学在国内的发展会产生一定的促进作用。希望这一新兴研究领域得到学界同行更多的关注。

（专题策划：梅剑华）

重复危机与实验哲学

Replication Crisis and Experimental Philosophy

梅剑华 /MEI Jianhua

（山西大学哲学社会学学院，山西太原，030006）
(School of Philosophy and Sociology, Shanxi University, Taiyuan, Shanxi, 030006)

摘　要：科学实验的基本要求是实验具有可重复性，实验哲学采取心理学实验办法进行哲学探究。因此实验哲学的实验也应该是可以重复的。近一些年来发生在认知科学、心理学领域的重复危机给心理学研究提出了很大的挑战，这也相应的对实验哲学提出了挑战。本文分析重复危机产生的原因，指出心理学实验和实验哲学实验的差异，进一步表明实验哲学能够回应重复危机所带来的威胁。

关键词：重复失败　基础比率错失　直接重复　间接重复

Abstract: The basic requirement of scientific experiments is that experiments are repeatable, and experimental philosophy adopts psychological experimental methods for philosophical inquiry. Therefore, the experiment of experimental philosophy should be repeatable. The replication crisis, in the fields of cognitive science and psychology in recent years have posed great challenges to psychological research, which has also posed similar challenges to experimental philosophy. This article analyzes the causes of repeated crises, points out

基金项目：高等学校学科创新引智计划“科技部 111 计划”（项目编号：D20021）。
收稿日期：2020 年 3 月 18 日
作者简介：梅剑华（1980–）男，湖北宜昌人，山西大学哲学社会学学院教授，研究方向为心智哲学、语言哲学、实验哲学与人工智能哲学。Email: jianhuamei@yahoo.com

the difference between psychological experiments and experimental philosophical experiments, and further shows that experimental philosophy can respond to the threats caused by repeated crises.

Key Words: Replication failure; Base rate fallacy; Direct replication; Indirect replication

中图分类号： N0 **文献标识码：** A DOI: 10.15994/j.1000–0763.2020.09.001

科学实验的一个基本要求是实验的可重复性，如果一个实验不可以重复，根据实验得出的结论就没有普遍性，不能算作科学实验。给定相同的条件，就会出现同样的事实。历史学家何兆武先生在《可能与现实:对历史学的若干反思》一书中，对这一问题有着清醒的认识：

> 历史是科学吗？回答是：历史学既是科学又不是科学，它比科学多了一点什么，又少了一点什么。我们所知道的历史，和自然科学所知道的历史有很大的不同。那就是自然科学所知道的事实，可以实验或实证，可是历史无法再做实验也无从实证。历史既然不能够重复，是一次性的，我们怎么样才能找到它的规律呢？我们普通所说的“规律”，都是在它重复了多次以后，我们才找到它的规律来。如果只是一次性的，那么我们怎么找规律？它有没有客观规律那种意义上的规律？如果没有的话，历史学家是凭什么理解历史的？自然科学家是凭借实验来理解自然现象，一次不对的话，可以再做。既然历史学家不能够做实验，那么他们怎么能够认识历史的真相？作为一个历史学家，了解历史的真相，就只有凭借自己的理解和想象。[1]

在何先生看来，历史不能成为严格的科学，是因为无法接受重复性的检验。通常有两种可重复性，一种是针对自然世界的；一种是针对人类社会的。自然科学的可重复性是相当精确的，通过控制、理论和预测来刻画。例如，伽俐略在比萨斜塔做的实验。所有与人相关的生命科学、医学、社会和行为科学均建立在统计的基础上，这种可重复性是概率的。基于统计推理的科学具有内在的缺陷：缺乏准确的理论作为预测，缺乏普遍的“真”、缺乏完善的实验控制手段。这些缺陷似乎都是研究者无法解决的问题。此外，基于统计的科学还有外在的缺陷，统计推断的可靠性差、研究者往往摆脱不了发表偏见。本文将首先引入在心理学领域的重复危机；其次，讨论重复危机对于实验哲学研究是否具有同样的意义，最后分析实验哲学中的重复失败案例。

一、重复失败与基础比率错失

重复危机（Replication Crisis, Replicability Crisis, Reproducibility Crisis）可有两层含义：第一是大规模的重复失败本身的危机；第二是相关科学领域必须把大规模重复失败作为事实接受下来所产生的科学信任危机。本文只关注第一种意义上的危机，尤其是这个意义上的重复失败。

在心理学、医学等依赖于样本测试的研究领域中，有很多原初实验无法为其他学者重复，成为孤例。实验哲学使用了心理学、认知科学大致类似的调查手段，依靠样本测试。如果重复验证失败对其它自然科学和社会科学是挑战，那么这对实验哲学也构成挑战。如果实验重复失败对其它自然科学和社会科学不构成威胁，那么对实验哲学也就不构成威胁。重复危机和重复失败相互关联，重复失败是重复危机的一种。通常出现重复危机的原因可能源自：不恰当的操作、不愿意发布负面的实验结果、数据欺骗等等。伯德（Alexander Bird）试图表明，即便重复失败频繁出现，那也不意味心理学、医学所使用的方法应该遭到质疑。[2]

科学研究试图表明存在一些特定的可重复效应，但一些重复实验却无法得出科学家所声称的效应，或者只有很小的样本才具有这种效应。有学者将这种情况归咎为实验者的科学实践存在问题，例如操作不当、实验设计存在缺陷、发表有隐瞒相反实验结果等等的问题。不难看出，上述缺陷原则上可以消解，并非严格意义上的重复失败。但有一类问题可以称为真正的重复失败问题：我们甚至可以预期会出现重复失败，这些失败是由于实验本身引起的，原则上找不到更好的实验来避免重复失败。为此也许可以采取更为积极、开放的态度去看待重复失败：第一、接受当前社会科学在本质上无法避免重复失败的事实，相应调

整对实验的评价；第二、在方法上，尽可能寻找可能为真的实验假设；第三、对实验的数据分析标准更加严格，例如，要求小于0.05的阿尔法值。

重复失败案例很多。例如在社会心理学研究中，有学者通过实验表明，如果在实验情景中提供一张双眼看着受试者的图片，受试者就会提高她们的社会合作行为，受试者更愿意给无人看守的箱子投掷硬币或零钱用来购买咖啡。[3]但更大规模的调查研究否证了这一效应。[4]既有的心理学实验告诉我们，小孩是天生的模仿者，但更大规模的调查研究否证了这一点。[5]

基础比率错失是重复失败的一个重要原因。就一般现象的某个特殊例示所做的推论经常会出现这种错误。基础比率错失并不是在计算概率时出现了计算错误，而是忽视了选择样本本身导致的偏差。例如，推理者关注到现象发生时的那些明显的证据，而忽视了即使独立于这些证据该现象也会发生的情况。比如说感冒的证据之一是头疼，但即便没有头疼你也依然可能感冒。这个没有头疼而患感冒的几率就是基础比率。尤其当现象（如渐冻症、飞机失事）少见，而证据和现象之间的关联不够紧密时，人们就会做出错误的推理。有一个著名的案例：在飞机场利用侧写工具扫描乘客的外表和行为，以此来判断他/她是否是恐怖分子。这种侧写工具的精度是95%：对于100个非恐怖主义分子，侧写工具的判断为：95个是非恐怖分子，5个是恐怖分子。对于100个恐怖分子，侧写工具的判断为：95个是恐怖分子，5个是非恐怖分子。设想这样一种情况：如果有一个乘客没有通过侧写测试。机器判断他为恐怖分子，那么他真的是恐怖分子的概率是多少？我们可能认为他有95%的概率是恐怖分子。实际情况并非如此，这个乘客是恐怖分子的概率极低。因为有大量的乘客作为统计的基数，恐怖分子是极少的。侧写的精度是基于对100个恐怖分子的测试，实际上一个机场里的几万人中恐怖分子的人数极少甚至没有。因此这个乘客更有可能是一个清白无辜的人，而非一个恐怖分子。这个关于恐怖分子的判断是机器的误差造成的。

卡塞尔的著名案例也反映了基础比率错失：实验调查向哈佛大学医学院的学生提出一个问题：如果有一个病人测试呈阳性（一种罕见的疾病），那么他实际上得病的概率是多少？学生同时被告知得病的几率为0.1％，阳性测试的精确率为95%。问卷调查结果表明，至少有一半的医学院学生给出的答案是95%，而事实上这种得病概率不到2%。[6]如果要检测的疾病非常罕见，那么检测受试者者呈阳性，就可能有两个原因：1）受试者没病，这个检测的精确率落在错误的5%中，所以呈阳性；2）受试者得病，这个检测的精确率落在正确的95%之中，所以呈阳性。实际上，更有可能的是，受试者没病。我们经常基于个人的经验做出推论而忽视某种现象在总体中发生的比率。例如，如果问一个人坐飞机危险还是坐火车危险。一般都会认为坐飞机危险。实际上飞机出事的概率要远远低于火车出事的概率。只是空难更容易见诸报端，让大家觉得坐飞机更危险。假设你正在乘坐飞机，突然发生剧烈的颠簸持续好几分钟，你甚至会涌现如下念头：完了，飞机要坠毁了。但是，如果调查大量的飞机颠簸，只有在非常少的案例中飞机颠簸之后坠毁，大量的案例则是经过持续颠簸后飞机正常飞行。具有飞行经验越多的人，心理素质越好。因为他的个人经验样本增大了,不会出现基础比率错失。通过持续的飞行体验，他会逐渐认识到飞机整体的出事率是相当低的。

让我们分析一下卡塞尔案例的实验方法。通常，检验假设使用随机对照实验、零假设检验来检测，其显著性水平在5%。实验者将样本分为两组：实验组和对照组。其中实验组进行干预，对照组不做干预。零假设指：对这两组的实验结果进行比对，假设它们之间不存在任何差异。如果这种统计结果p值小于0.05，就应该拒斥零假设。显著性水平为0.05表明，如果接受零假设为真，那么就会有5%的错误率去拒斥零假设。这种假设检验的方法在这方面有95%的准确率，接受错误率为5%。卡塞尔案例是一种非典型情况，实验现象（实际病例）极为罕见，因此，我们如果就这一病例提出假设，那么假设为真的概率只有2%。这和测试恐怖分子案例是一样的道理，如果掌握总体，来对具体现象做统计，至少会区分出发病率极高和发病率极低两种情况。在日常生活中，我们更容易高估一些案例的发生率，例如空难、罕见的癌症等等。还存在一种典型情况，实验现象（实际病例）比较普遍，假设为真的概率是10%，测试方法的精确率为95%。如果测试为真，实验假设为真的概率是多少呢？计算可以知道假设为真的概率为68%。对于一个学科来说，如果面对一

些现象，提出的假说足够多，那么由于显著性差异检验的5%容错率存在，就会有一定数量的假说被实验统计得到为真，其中大部分都是假阳性（false positive）的结果。如果以上成立，那么即便某个心理学实验关于一个假说得到了一个阳性结果（positive result），它很大概率是假阳性。因此不能被重复是注定的。这个考虑不会对实验哲学有所冲击，在实验哲学家所设计的一个具体的调查实验中，提出的假说并非足够多，而是非常少的。假设某一个调查问卷，涉及两个选择支甲和乙。那么，我们就会预测：选择甲的受试者具有a理论直觉，而选择乙的受试者具有b理论直觉。反驳者有可能提出其它因素导致受试者选择甲或者乙，但这种反驳是有限的，可以逐一加以排除。当然这需要进一步做实验调查，而非一次性调查。

非典型情况和典型情况二者都有可能发生重复失败情形，根本原因在于基础比例的问题。这里有两种类型的精准：第一种是当x如此的时候，x如此的准确率；第二种是当x并非如此的时候，x并非如此的准确率。研究置信水平，相应存在两种类型的不准确性：当x实际如此时，判断x并非如此，这称之为Ⅱ型错误（false negatives假阴性）；当x实际并非如此时，判断x如此这般，这称之为Ⅰ型错误（false positives假阳性）。重复失败也许是因为有些现象的重复率本来就很低。

二、直接重复和概念重复

菲斯特（Uljana Feest）认为，重复危机对于实验研究来说并不是一个严重的问题。[7]也许更重要的是，在研究中所遇到的概念的、实质的问题。重复性并不如通常想象的那样重要。重复性危机争论中涉及到两个相互关联的问题：1）为什么没有更多的实验重复？ 2）为什么60%的研究不能被重复？统计分析中，有一个广泛存在的p–hacking现象：科研人员在分析数据时不断尝试计算统计数值，直到得出p值 < 0.05才会发表结果，这种操作导致后继者很难重复实验。熟悉零假设检验的人都知道，研究者总是在试图发表否定零假设的研究，来证明自己假设的正确性。为了达到费希尔（Ronald Fisher）人为设定的$p < 0.05$的显著性标准，很多研究者可能夸大了研究中的效应值（effect size）。导致p–hacking的常见行为包括：因为$p < 0.05$而放弃实验数据的收集；测量一大堆因变量，再根据p值选择性地报告因变量测试结果；根据p值删掉异常的原始数据（outlier）；根据p值决定如何处理实验组；在实验进行过程中篡改数据等等。上述行为实质上都是一种学术造假，隐藏在一大堆数据背后的造假导致了后续的重复实验失败。

实验设计与我们对概念和事实的理解不可分割。我们对问题本身的预先理解会影响实验的结果。我们可以把实验重复分为两种类型：直接重复和概念重复。直接重复是原样重复；概念重复是基于基本预设框架的重复。二者细节有所不同。菲斯特考察已有的实验，发现大部分既非直接重复也非概念重复。“重复”的字面意思是指：原初的实验和重复的实验是同一个实验。严格来讲，很难重复同一个实验：最起码来说，二者的时间参数就不一样。如果把时间参数考虑在内，那么，重复实验就是原则上不可能的。套用个人同一问题中“数的同一”（numerical）和“质的同一”（qualitative）的区分，大多数实验并不是追求“数”的同一，而是追求“质”的同一，只要高度相似就可以了。所以，不可能实现完全一样的重复实验，只存在尽可能高度接近原初实验的重复实验。这种实验重复可以称之为直接重复（direct replication）。即便如此，“相似/接近”也对实验可重复造成麻烦。在实验环境中，很多明确描述的实验条件已经预设了概念的、实质性的一些假定，但这些预设并未明确表述出来。因此，这些隐含的特征如何在另外一个实验中完全复制就成了一个问题。例如，假定研究者关心“听莫扎特音乐是否能够促进小孩智力”，他们就需要设计一个实验，把莫扎特的一段音乐作为自变量，把测试出来的智力分数作为因变量，研究结果称之为莫扎特效应。研究者需要在自变量和因变量二者之间建立因果联系。一旦研究者认为实验的结果是关于智力的，那么就隐含预设了研究者的测试是测试智力的。这些判断都通过对自变量和因变量的选择而实质依赖于非经验的概念假设。这不单单是实验的设计和操作问题。研究者如果在特定的操作程序下进行实验，那么，去判断实验成功与否就不能避免概念预设问题。也许有人会回应说，重复并不需要研究描述之下的因果关系，只需要单纯模仿重复即可，不需要对干涉或操控做出特定的承诺。但即便如此，“个人判断”也会不可避免地干扰实验。因为，有很多原初实验的辅助性假设并没

有被完全明确地表述在实验设计报告中。新的实验里，研究者很可能有意或无意地增加或者减少了相关未被明述出来的因素。如果完全一模一样的重复实验不可能的话，那么基于相似性的重复，就总是有可能错解相似性。例如，在原初实验中隐含假定了实验室内的颜色和实验无关，那么就可能会在重复实验时也会忽视这一原因，但实际上很可能颜色在原初实验中是起作用的。这个实验者所做的个体化判断使得实验在认知上变得不太确定。在选择变元时，就出现了概念化范围的问题。例如，我如何来描述听莫扎特音乐这个刺激相关联的特征呢？是莫扎特这首曲子本身、大调还是别的什么？这都依赖于实验者如何描述、如何概念化。如果我把曲子作为相关特征，我就会把其它莫扎特曲子作为对照；如果我把曲调作为相关特征，我就会把情绪作为对照。实际上，这种选取并没有严格正确性的标准。概念化问题不可能从实验研究中清除，概念问题与经验问题在实验设计中同时存在。

直接重复试图尽最大可能复制原初实验，概念化重复则试图用不同的方法去解决同一个问题。概念化重复有其自身的优点，可以把实验的结果一般化。我们不妨对一般化做进一步区分：1）针对这个实验可以做出什么类型的推论？ 2）这个实验的结论也适用于实验室之外的情景吗？我们可以用内部推论和外部推论来简单概括。因为混杂共因的存在导致认知的不确定性，从而使得内部推论不成立。以莫扎特音乐实验为例，如果部分受试群体早已受过音乐教育，这就会影响测试智力的结果。而且，由于不知道哪一种概念化（是选择莫扎特还是选择调性）是正确的，也会导致推理错误，这是寻找操控变元的错误，可以为概念化重复是不可能的这一问题提供回应思路。因此，概念化重复失败几乎无法避免。假如我想比较两个针对同一个问题的不同实验设计的结果，那么我实际上已经预先假设了这两种操作事实上具有相同的概念范围，但这恰恰是问题所在。

杜伊恩（Stéphane Doyen）等指出，在没有直接重复的情况下，概念重复的问题变成了这样一个问题，即根本就不存在“概念上重复失败”这样的事情。原因在于，使用不同操作而没有找到相同“效果”，可以归因为方法的差异，而不是原初实验结果的不可靠性或不稳定性。一般只会发表成功的概念重复实验，而不成功的概念重复实验基本都被完全忽视了，但它所依赖的基础却没有受到质疑。

总之，无论是直接重复还是概念重复，都会面临失败，而这种失败可以事先预期。或许，我们不得不接受这样一个事实：重复失败是心理学研究面临的一个事实，而非心理学研究的错误所导致。那么接下来的问题就是，这种重复失败对心理学研究是一个威胁吗？菲斯特分了三步进行了论证：第一、直接重复排除了随机误差，但回避不了系统误差。如果接受在直接重复中不可避免地引入了个体化判断，那么系统误差就总是存在的。第二、如果我们严肃对待心理学研究中的实验，就会发现，相当多的研究结果都是不确定的，但这并不必然构成危机。例如，围绕莫扎特效应的一些实验研究，实际上很多都是概念工作，很多可能的错误都源自对概念范围划分的不明确。第三、重复实验产生的问题描述为“重复危机”并不妥当。相反，我们重复实验里看到的是这样一个情形，即科学家们在面对高度的认知不确定性和概念开放性时，探索一种假想效果的经验轮廓。我们可以说，关于当代心理学研究存在“信心的危机”，但这并不是重复危机。

三、实验哲学中的重复危机问题

有学者指出，心理学和实验哲学中重复危机的根本原因在于，这类实验都是关于人的，尤其是关于人的心理层面而非物理层面的。“人心难测”，人在不同环境下的心理状态在根本上的不同，造成了每次实验结果的不同。物理学实验不会导致这种问题，通常，物理学实验的对象本身不会影响实验者的研究。从实践和理论的区分来看，在实践层面，操作者人为引进了一些干扰因素，造成重复失败；在理论层面，由于每次实验的时空非均质化，造成了任何实验都面临失败的可能性。在普遍失败的基础上，心理与物理的区分也会造成心理学实验的失败率要高于物理学实验的失败率。但是，物理学实验亦需要区别对待。例如，在微观领域，量子现象的存在，造成每次的实验都会出现差异，这种干扰现象也会导致重复失败。但心理学实验和物理学实验的不同在于，心理学的实验依赖于庞大的样本，而不是单个对象的实验。心理学实验中受试者个体的变化和微观层面的变化可能不会导致样本特征的整体改变，这是因为，要造成样本整体的改变，就意味着每一个体都

在某一方面发生同样的变化，具有统计上的显著性。如果真的具有这样一个改变，那正是调查研究所需要发现的。通过多次重复实验，就能搞清楚这种系统改变。但一般来说，个体的变化不会影响到样本。

心理学认知类杂志，例如《实验心理学杂志：学习、记忆与认知》，报告的实验可重复率为48％–53%，社会认知类杂志，如《人格与社会心理学杂志》，报告的实验重复率为23％–29%。实验哲学既关注认知心理问题，也关注社会心理问题。由此可以粗略猜测，实验哲学的重复率在二者之间，大概为35%，但这个结论可靠吗？我们考察一下实验哲学家对这一重复率的调查情况。科瓦（Florian Cova）等从OSC（Open Science Collaboration）计划中建立了实验哲学重复计划，包括20个实验重复团队的40位研究者，跨越8个国家，以评价实验哲学的重复率。[8]科瓦等用三种方法检测重复实验成败与否：（1）重复结果是否统计显著？通常统计显著指p值小于.05，但p值不应该是唯一的衡量标准。（2）重复实验团队的主观评价。尽管这种评价不是非常可靠，但如果p值并非唯一衡量标准，那么加入一些研究者关于实验设计的思考，从更为宽广的视角来看待实验所涉及的问题，大有裨益。（3）对比原始实验和重复实验的效应量（effect size）。关键不在于是否存在这种效应，而是这种效应的大小。同一个重复实验，以p值的标准看是成功的，但从理论角度来看则未必。科瓦等的实验中，研究者选取从2003年到2015年间的40个实验哲学中的实验：行动理论8个，美学1个，因果4个，认识论5个，自由意志8个，道德心理学8个，语言哲学1个，心灵哲学2个，未分类3个。在挑选样本时，为了得到一个合理的样本，采取了两种办法：第一、抽取每年引用率最高的实验；第二、在每年的剩余实验中，再随机抽取一些实验。

重复结果对照如下：实验者的主观评价：调查实验重复者对重复情况的判断。在40个实验中，有31个被认为重复成功，重复成功率为77.5%。统计检验情况，原始实验的均值N是215.1（SD=542.3），重复实验的均值206.3（SD=131.8）。[9]当p值小于0.05，实验就被认为是重复成功。有3个实验得出了零结果被排除在外。结果表明，重复成功率为78.4%。比较效应量：撇开细节，统计结论显示二者之间的效应量没有差别。[9]总体的结论是：实验的可重复性较高。与心理学重复相比，心理学的重复成功率在36.1％–47.4%之间，实验哲学的重复成功率大概要超过70%。实验哲学能够取得较高的重复，主要归因于以下几个方面的原因：

第一个原因：实验哲学的高重复率源于较大的效应量。OSC 报告的均值r的效应量是（M=0.403，SD=0.188），这实际上要比我们平均原始值r 的效应量要高（M=0.38，SD=0.16）。这种初期评价可能受到发表偏见、相对较小的样本以及其他因素的影响。

第二个原因：实验哲学实验要比心理学实验更容易获取大的样本，而且耗费较少。因为很多实验哲学研究的实验调查都较为简单。首先，较为简单的研究可以很容易获取大的样本，从而得到比较可靠的结果；其次，实验哲学所设计的实验对于研究者来说，所耗费的时间和资源是相对较少的。研究者花费的精力越多，就越希望发表自己的成果，这可能造就了一些很成问题的实验研究。这种希望发表自己成果的欲望也可能增大类型Ⅰ错误。为了验证这一推测，科瓦等让研究者对40个实验哲学中的实验难度进行评级：分数为从0到2，得1分指：所进行的实验不仅仅是网络调查，可能还需要一定的实验条件。得2分指：受试人群选取困难，例如要选不同文化传统中的群体。 评级结果显示：没有实验被评价为2，36个实验的分数为0，4个为1。这说明与心理学实验相比，实验哲学的实验设计是简单的，而且难度较大的实验重复率，要明显低于难度较低的实验重复率。例如，高难度实验的重复成功率为50%，而低难度实验的重复成功率为80.6%。当从心理学开放合作平台中抽取99个实验作对比时，结果发现，12个实验被评价为2，17个实验的分数为零，70个实验的分数1。这表明心理学实验的确要比实验哲学中的实验难。但在心理学实验中，难易程度本身并不影响实验的重复成功率。容易的（0）实验成功率为43.8%；中等程度（1）的实验成功率为38.2%；困难的（2）实验成功率为36.4%。这表明：在心理学实验中，可能存在其他影响因素。

第三个原因：效应类别。研究者所选择的实验一般属于以下四类：一、搜集观察到的数据，如语料。二、基于内容的实验。如改变实验内容中的条件，看受试者如何反应。在诺布效应实验中，把“对环境有害”改为“对环境有益”，来考察受

试者回答的差异。三、基于环境的实验。因为实验本身框架所引起的受试者回答的差异，如故事场景的人称差异（从第一人称改为第三人称）、给受试者以认知负荷等等。四、人口效应：参与者的文化差异、个性特征导致回答的差异。根据已有的观察，可以提出两个假设：其一、大部分实验都属于第二类，通过改变内容中的条件来获取受试者的数据差异。其二、第二类的实验结果要比第三、四类更稳定。例如，框架效应依赖于受试者的注意力以及是否察觉到框架的干预。而人口效应可能会被群体内的实验结果所弱化。譬如指称直觉测试中，所得结论是：东方人倾向描述直觉；西方人倾向因果历史直觉。这就是人口效应。但如果进行群体内差异检测，在东方人内部也发现系统差异，一部分人倾向描述直觉，另一部分倾向因果历史直觉。这就弱化了人口效应的影响。为了验证这一猜想，科瓦等对已有的40个实验进行分类：其中一个实验属于第一类；31个实验属于第二类；4个实验属于第三类，4个实验属于第4类。这个分类结果表明，实验哲学研究大多是基于内容研究的，主要关注在给定故事里的内容变化（刺激）如何影响受试者的回答（反应）。统计结果表明：基于内容研究的实验重复率，要高于基于框架和人口效应的实验重复率。然而，在心理学实验中，基于内容研究实验的重要程度，要低于实验哲学。

实验哲学中也存在实验偏见问题：麦希瑞（Edouard Machery）等分析既有实验哲学材料是否具有证据价值（evidential value）。[11]实验哲学家认为，这些研究为哲学问题提供了洞见，另一些哲学家和心理学家认为这些研究缺乏证据价值。哲学家意见分野的重要原因之一就是，实验存在选择性偏见，只选择那些具有统计显著的结果发表，并且存在p–hacking（增加在显著水平以下的p值）。麦希瑞搜集了365篇文章进行p–曲线分析。结果表明：这些文献结论都具有证据价值。P–曲线指：对于实验的全集统计显著性p值的一个分布。P曲线偏离均匀分布显示了，是否是因为选择偏差或p–hacking。统计显著性要比通常所认为的弱一些。诺布（Joshua Knobe）等人组建了实验哲学重复计划网站（Experimental Philosophy Replication Page）。①在这个网站上，重复成功率为57.6%。虽然这个重复成功率比预测的要高，但还是说明实验哲学的结果并非完全可靠。不过，这种对重复失败的考虑是有问题的，因为人们愿意去重复的实验大部分都是一些让人觉得新奇的、有趣的实验，而那些常识化结论的实验没有人愿意去重复。这背后存在着所谓的重复选择偏见，或者样本选择误差。

实验哲学和心理学重复成功的差异可能在于：实验哲学主要集中于一些非常稳健的效应，而传统心理学则更关注一些细微效应，尤其是外部环境影响的效应。简单来说，实验哲学更关注内部效应，而心理学则更关注外部效应。在我看来，这恰恰可能是因为实验哲学处于早期阶段，而心理学处于成熟阶段所引起的后果。任何一门学科在开创之初都会聚焦于内部问题，但是随着研究的深入，内部问题研究的越来越系统，就不得不转向一些外部性问题。对照的结果实际上就是：正在发展的实验哲学与比较成熟的心理学的对照。有一个对实验哲学的批评与此相关，即实验哲学得到的结果通过概念分析和内省反思也可以得到，因此，实验哲学不过是重复了传统哲学的结论，了无新意。 这个批评源于很多实验哲学的早期工作都深深植根于哲学传统，通过实验哲学所获得的那些结果，足够稳健，使得哲学家通常可以通过内省反思获得。但随着实验哲学的深入发展，实验研究所获得的结果就有可能和通过内省反思得出的结论相反。原因在于，所谓通过内省反思获得，归根到底也是藉由和世界打交道形成的经验所塑造出来的内省反思，但每个人的经验都是高度局限的，属于样本中的一个个例。高度个人化的思考不能反映人类认识世界的系统现象，作为认知者本人，察觉不到认知产生的系统差异也是很正常的。实验哲学的深入研究有利于纠正概念分析的片面与局限。

心理学和实验哲学重复失败的差异还在于学术共同体文化。廖显祎指出，哲学家对于诸如“什么可以作为论断的证据的方法论问题”更为敏感，对重复实验也更为宽容。[11]与传统心理学家相比，实验哲学家在发表实验结果上的压力要小于心理学家。发表有数据分析的论文不是哲学家唯一的出路，但发表有数据分析的论文是心理学家几乎唯一的选择。这种差异性给心理学家造成了收集、

①https://sites.google.com/site/thexphireplicabilityproject/home.

分析数据的压力，亦造成所谓的发表偏见。有人指责实验哲学不过是心理学，而且是坏的心理学。心理学所具有的问题，实验哲学也有（重复危机）；心理学没有的问题，实验哲学亦有（如结论无法推广）。所以，实验哲学被认为既是糟糕的哲学，又是糟糕的心理学。假如换个角度看，也许实验哲学不仅是好的心理学，还是好的哲学。因为实验哲学既具有心理学所缺乏的概念分析视角，又具有传统哲学所缺乏的经验视角，既注重概念分析又注重经验探索是实验哲学所具备的优势。

结　　论

在重复危机的讨论中，我们应该回答这样一个总结性的问题：在心理学中，重复危机发生的原因有哪几个？其中有哪些是实验哲学可以避免的，哪些又是不可避免的？在心理学中，重复危机发生的原因可以总结为三类：第一、基础比率错失，没有考虑到样本本身的偏差；第二、发表偏见，即只发表对结论有利的研究结论；第三、直接重复和概念重复本身引起的重复困难问题。其中，前两个原因是能够通过实验者调整实验设计和数据分析而加以避免的，也是大量存在重复危机的原因。第三类原因却是实验者不能通过调整实验设计和数据分析来避免的，严格来讲不能算作重复危机的真正原因。实验哲学是运用心理学调查方法研究传统哲学问题的一个跨学科研究领域。一方面，我们可以说实验哲学中的实验就是心理学实验，因此心理学和实验哲学所面临的重复危机是一样的；另一方面，由于实验哲学本身研究对象的差异，又有所区别。例如，实验哲学的实验数据发表压力，小于心理学的实验数据发表压力，因此发表偏见产生的问题要小一些；实验哲学家对直接重复和概念重复的问题，有比较高度的自觉意识，会主动寻求解决之道；实验哲学尚处于成长期，相对容易发表有明显结果的研究等等。

总之，从实验哲学研究来看，我们可以提前了解实验中的重复失败现象，避免基础比率错失和滥用数据的问题。虽然不可能进行直接重复和概念重复，但这并不意味着我们不能用实验结果拓展既有的研究思路。既有的研究也表明，实验哲学的重复成功率要高于心理学的重复成功率，我们可以在具体的实验中消除实践引起的误差，给心理学和实验哲学的实验重复失败找出一个合理的解释。重复失败以及因此而引起的重复危机，不会给实验哲学带来严重威胁。

［参 考 文 献］

[1] 何兆武. 可能与现实：对历史学的若干反思[M] 北京：北京大学出版社，2017, 83.

[2] Bird, A. 'Understanding the Replication Crisis as a Base Rate Fallacy'[J]. *The British Journal for the Philosophy of Science*, 2018. https://doi.org/10.1093/bjps/axy051.

[3] Bateson, M., Nettle, D., Roberts, G. 'Cues of Being Watched Enhance Cooperation in a Real–world Setting'[J]. *Biology Letters*, 2006, (2): 412–414.

[4] Carbon, C., Hesslinger, V. M. 'Cues–of–being–watched Paradigm Revisited'[J]. *Swiss Journal of Psychology*, 2011, (70): 203–210.

[5] Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy–Costantini, S., Davis, J., Slaughter, V. 'Comprehensive Longitudinal Study Challenges the Existence of Neonatal Imitation in Humans'[J]. *Current Biology*, 2016, (26): 1334–1338.

[6] Casscells, W., Schoenberger, A., Graboys, T. B. *Interpretation by Physicians of Clinical Laboratory Results*[J]. *New England Journal of Medicine*, 1978, 299(18): 999–1001.

[7] Uljana, F. 'Why Replication Is Overrated'[J]. *Philosophy of Science*, 2019, 86(5): 895–905.

[8] Doyen, S., Klein, O., Simons, D., Cleere–mans, A. 'On the other Side of the Mirror: Priming in Cognitive and Social Psychology'[J]. *Social Cognition*, 2014, (32): 12–32.

[9] Cova, F., Strickland, B., Abatista, A. 'Estimating the Reproducibility of Experimental Philosophy'[J]. *Review of Philosophy and Psychology*, 2018, (10): 1–36.

[10] Stuart, M., Colaço, D., Machery, E., 'P-curving x-phi: Does Experimental Philosophy Have Evidential Value?'[J].*Analysis*, 2019, 79(4): 669–684.

[11] Liao, S. 'The State of Reproducibility in Experimental Philosophy'[EB/OL]. https://philosophycommons.typepad.com/xphi/2015/06/the–state–of–reproducibility–in–experimental–philosophy.html. 2020–1–15.

［责任编辑　王巍　谭笑］