5th BEIJING ANALYTIC PHILOSOPHY CONFERENCE

28 April 2018

JUSTIN SYTSMA
(Victoria University)

"Are Religious Philosophers Less Analytic?"

COMMENTS

MEI JIANHUA
(Capital Normal University)

Venue Room 500
Renwen Building
Renmin University

Language English

Organizers Huang Yudi
Li Yongna
Liu Xiaoli
Liu Chang
Tao Stein
Tian Jie
Daniel Lim

CONFERENCE SCHEDULE

9:20 – 9:30	**Introduction**
9:30 – 10:10	论 Dretske 对直接/ 间接知觉之别的批评 (1-6) Hua Haiming (SunYat-sen) Comments: Liu Chang (Renmin)
10:15 – 10:55	*Mixed-Effects Modeling and Non-Reductive Explanation* (7-31) Fang Wei (Tongji) Comments: Stefan Petkov (Beijing Normal)
11:00 – 11:40	*Two Ways of Mental State Attributions* (32-46) Wu Minyang (Renmin) Comments: Justin Sytsma (Victoria)
11:45 – 12:25	*Subjective Beliefs in Outcome Probability and Moral Decision in Moral Dilemmas* (47-53) Song Fei (Hong Kong) Comments: Li Yongna (Renmin)
12:30	**Group Photo**
12:45 – 1:45	**Lunch Break**
1:45 – 3:00	**Keynote Address**
	Are Religious Philosophers Less Analytic? Justin Sytsma (Victoria University) Comments: Mei Jianhua (Capital Normal University)
3:00 – 3:10	**Essay Prizes & Closing Comments**

论 **Dretske** 对直接/间接知觉之别的批评

花海铭

摘要：直接/间接知觉之别是知觉哲学的核心主题之一。根据 Snowdon 的看法，有三种对待这个区别的态度，而 Dretske 是其中一种态度的代表人物。Dretske 对这个区别采取批评的态度。他论证说，最流行的三种划出这个区别的方法都不能达成人们最初引入这个区别的目的。但是，笔者企图论证：如果有一种融贯的划分方法可以捕捉到“直接知觉”和“间接知觉”这对词语在哲学圈当中的用法的话，那么 Dretske 的批评并没有触及到这种划分方法。笔者的理据是，他所批评的那三种划分方法当中的第一种和第三种违反了一条包含这对词语的分析真理，第二种则没有尊重错觉论证的修正性。

关键词：直接知觉；间接知觉；错觉论证；感觉材料

Abstract: The distinction between direct and indirect perception is one of the core themes in philosophy of perception. According to Snowdon, there are three attitudes towards this distinction, and Dretske represents one of them. The attitude he takes is a critical one. He argued that even the three most popular methods of drawing this distinction failed to achieve what they were intended to achieve. However, I attempt to show that if there is a method which captures the philosophical usage of the pair of words “direct perception” and “indirect perception” coherently, then Dretske's criticism hasn’t damaged this method. This is because the first and third of the three methods he criticized violate an analytic truth that contains the pair of words, and the second one doesn’t respect the fact that the argument from illusion is a revisionary argument.

1，引言

直接/间接知觉之别是知觉哲学当中一个相当重要的区分。当我们考虑错觉论证的时候，这一点就会变得非常明显，因为错觉论证企图证明：我们所直接知觉到的从来都不是物理对象，而是感觉材料。根据 Snowdon（1992: 50-51）的看法，有三种对待这个区分的态度。除了他本人所代表的那一种态度以外，其余的两种态度都是负面的。其中，以奥斯丁（1962: 15）为代表的态度认为，哲学家歪曲了“直接看见”和“间接看见”这类词语对子的日常用法，最终导致它们没有意义；以Dretske（1969: 62-75）为代表的态度认为，这个区分一旦经过澄清之后，我们就会发现它并不能达成人们最初引入它的目的。本文的工作是要论证，如果有一种融贯的划分方法可以捕捉到“直接知觉”和“间接知觉”这对词语在哲学圈当中的用法的话，那么 Dretske 的批评并没有触及到这种划分方法。Dretske 自称批评了划分直接/间接知觉之别的三种最流行的方法，下文逐一讨论。

2，第一种划分方法

第一种方法主张：主体直接看见 x，当且仅当，（主体看见 x，而且如果主体相信 x 具有或没有某个视觉性质，那么这个信念不可能是错的）。Dretske 指出，如果我们这样子刻画直接知觉，那么没有什么东西是我们直接知觉到的。“凡是我们看见的，凡是我们视觉上觉察到的，都是间接地看见的”（Dretske 1969: 64）。首先，日常的物理对象自然被排除在外了，因为我有可能误以为我面前的杯子是咖啡色的，虽然它实际上是黄色的。然而，就连那些常常被哲学家当作直接知觉的对象的事物（例如后像、幻觉里的匕首、海市蜃楼、眼冒的金星）也被排除在外了。这是因为，在下述情况下，我甚至会弄错一个后像的颜色：我看见了一个紫色的后像 a，a 在我的视野当中的位置使我误以为它是沙发上的污迹，而我碰巧知道这张沙发有一片蓝色的污迹，于是我误以为 a 仅仅是因为奇怪的照明而显得是紫色的，其实它是蓝色的。由此可见，虽然引入

直接/间接知觉之别的初衷之一是想确保感觉材料是我们所直接知觉的，但这个划分方法并不能达成这个目的。

笔者认为，如果 Dretske 对这第一种划分法的批评是成立的，那么错觉论证是不可靠的（unsound）。为了看出这一点，我简略地分行表述错觉论证的常见形态[1]：

P1，在错觉当中，一个物理对象 a 看起来是 P，但其实不是 P。

P2，如果 a 看起来是 P，那么主体直接知觉到了一个是 P 的对象 b。

P3，附近没有任何相干的（relevant）物理对象是 P。

C1，在错觉当中，主体直接知觉到了一个非物理的对象。

P4，如果（C1）为真，但是主体在正常的情况下所直接知觉到的对象却是物理的，那么错觉和正常的知觉就不可能有时候是无法分辨的。

P5，错觉和正常的知觉有时候是无法分辨的。

C2，主体在正常的情况下所直接知觉到的对象也是非物理的。

C3，不论是在错觉当中还是在正常的情况下，主体所直接知觉到的都不是物理对象。

易见，上述的（P2）为假。这是因为，虽然事物总是看起来具有某些性质，但是没有什么东西是我们所直接知觉的。这个结果固然是间接实在论者不愿意看到的，但直接实在论者其实也并没有从中获得任何优势，因为根据这第一种划分方法，我们也不能直接知觉到物理对象。更糟糕的是，笔者发现 Dretske 所给出的这第一个划分方法违背了下面这条很可信的规则（ID）：

ID. $\forall x \exists y$ (主体间接知觉到 x → 主体直接知觉到 y)

既然凡是我们知觉到的都是间接知觉到的，而没有任何东西是我们直接知觉到的，那么（ID）为假。但是，这条规则是使用直接/间接知觉之别的哲学家都应该赞同

[1] 笔者此处的表述受惠于 Huemer (2001; 2011) 以及 Snowdon (1992)。

的[2]。我们甚至可以说（ID）是一条分析真理——如果“直接知觉”和“间接知觉”这对词语在哲学文献当中的用法有任何意义的话，它们的意义必须使得（ID）为真。由此可见，这第一种划分方法并没有捕捉到这对词语在哲学圈当中的用法。

3，第二种划分方法

目光转向 Dretske 所考虑的第二种划分方法：主体直接看见 x，当且仅当，（主体看见 x，而且，必然地，如果 x 看起来是 P，那么 x 就是 P）。Dretske 承认，这个区分一方面把物理对象排除在直接知觉的范围以外，另一方面把感觉材料的常见候选人（比如后像）纳入到该范围之内。毕竟，一个看起来是圆形的后像不可能不是圆形的。但是，Dretske 批评说，这个划分方法并没有带来大家所预期的知识论意义[3]。并不会因为我所看见的对象是感觉材料，我就更容易发现该对象有哪些性质。为了说明这一点，Dretske 打了一个比方。假设有这么一群人，姑妄称之为“开放人”，他们喜怒形于色，从来都不掩饰内心的真情实感。但是，并不会因为我碰上的是一个开放人，我就更容易察觉这个人的内心感受，因为问题只会转变为我如何分辨他是不是开放人。

笔者认为，Dretske 对这第二种划分方法的讨论有两点破绽。第一，这第二种划分方法也没有捕捉到那对词语在哲学圈当中的用法。这是因为，作为错觉论证的结论的其中一部分[4]，“我们不曾直接知觉到物理对象”是一个修正性的论断[5]，而不是一个无聊的论断——这个论断要求我们修改常识的观点。但是，按照目前所讨论的这种划分方法，这个论断就会变成一个无聊的真理，它不需要我们修改常识的观点。毕竟，没有一个普通人会否认物理对象看起来的样子至少有时候是误导人的，常识并不否认错觉的存在。第二，开放人的比方是误导人的。虽然要分辨一个人是不是开放人是一

[2] 例如，罗素 (2001: 3-4) 下面这段话就正好暗示了直接知觉是间接知觉的必要条件：“实在的桌子假如确实存在的话，也不是被我们所直接认识的，而必定是从我们所直接认识的东西中得出的一种推论。”

[3] 关于这一点，可参照奥斯丁 (1962: 104-131) 对所谓“不可被纠正者”(the incorrigible) 的批评。

[4] 注意，这个论断只是错觉论证的结论的负面部分。结论的另一部分是正面的：我们所直接知觉到的是感觉材料。

[5] 关于该论断的修正性，可参见 Snowdon (1992: 52)。

件不容易的事情，但是我们凭生活经验可以很容易地分辨出感觉材料，所以用前者去类比后者是不恰当的。具体说来，由于过往的生活经验，我们已经熟悉了下面这类规律：困意来袭的时候，看东西会看成双的；头部被重击之后会眼冒金星；如果烈日当空的时候看太阳，我们会留下后像等等。这些规律可以帮助我们预测感觉材料的出现。毕竟，没有一个成年人犯困的时候会认真相信自己看见了两支笔。总言之，关于感觉材料的原初知识（proto-knowledge）并没有这么难获得[6]。

4，第三种划分方法

最后，我们考察 Dretske 所讨论的第三种划分方法：主体直接看见 x，仅当，主体看见 x 的时候 x 必然存在。Dretske 承认，按照这种刻画，我们不曾直接知觉到物理对象。这是因为，光线从杯子反射到我们的眼睛，再经过某些大脑事件，再到那个被我们称为“看见一只杯子”的事件发生，这个过程需要一定的时间（即使极短）。这么一来，当最后那个事件发生的时候，那个杯子总是有可能已经不存在了。幸运的话，我们充其量也只是间接地知觉到杯子。Dretske 对这个刻画的批评是，“我们仅仅间接知觉到这个杯子”并不保证“我们直接知觉到了某个东西”。他让我们考虑一个类比：让我们规定，x 直接收到 y 的信息，仅当，x 收到信息的时候 y 必然存在。诚然，当我们收到某人的信息的时候，这个人总是有可能已经不存在了，所以我们仅仅间接地收到这个人的信息。但是，难道这就保证一定有一个第三者，我是从他那里直接收到信息么？笔者认为，Dretske 的论敌大可以爽快地回应说：“诚然，前述的保证关系并不成立。但是这并不要紧，因为感觉材料在我们知觉到它们的时候总是必然存在的，而这一点并不是由间接知觉和直接知觉之间的逻辑关系所造成的，而是由感觉材料的本性所造成的”。此外，Dretske 的批评显示出，这第三种划分方法否定了（ID）的分析性。换言之，Dretske 的批评暴露出，即使（ID）是真的，它也不是分

[6] 关于原初知识，可参见 Dretske (1969: 93-112)。

析地真。这就表明，这第三种划分方法也没有捕捉到“直接知觉”和“间接知觉”这对词语在哲学圈当中的用法，因为这种方法没能保留（ID）的分析性。

5，总结

在以上的讨论中，笔者检视了 Dretske 对直接/间接知觉之别的三种划分方法的批评。过程中，笔者论证了，如果有一种融贯的划分方法可以捕捉到“直接知觉”和“间接知觉”这对词语在哲学圈当中的用法的话，那么 Dretske 的批评并没有触及到这种划分方法。笔者的理据是，这三种划分方法当中的第一种会使得（ID）为假，第二种则没有尊重错觉论证的修正性，第三种则未能保留（ID）的分析性。

参考文献

Austin, J. L. (1962) *Sense and Sensibilia*, Oxford: Oxford University Press.

Dretske, F. (1969) *Seeing and Knowing*, London: Routledge & Kegan Paul.

Huemer, M. (2001) *Skepticism and the Veil of Perception*, Lanham, Md.: Rowman & Littlefield.

Huemer, M. (2011) “Sense-data.” In: Zalta, E. N. (ed) *The Stanford Encyclopedia of Philosophy*, Spring 2011 edn. http://plato.stanford.edu/archives/spr2011/entries/sense-data/ Russell, B. (2001) *The Problems of Philosophy*, Oxford: Oxford University Press.

Snowdon, P. F. (1992) “How to Interpret ‘Direct Perception’.” In: Crane, T. (ed) *The Contents of Experience*. Cambridge: Cambridge University Press, pp. 48–78.

Mixed-Effects Modeling and Non-Reductive Explanation

Fang Wei

Abstract: This essay considers a mixed-effects modeling practice and its implications for the philosophical debate surrounding reductive explanation. Mixed-effects modeling is a species of the multilevel modeling practice, where a single model incorporates simultaneously two (or even more) levels of explanatory variables to explain a phenomenon of interest. I argue that this practice makes the position of explanatory reductionism held by many philosophers untenable, because it violates two central tenets of explanatory reductionism: single level preference and lower-level obsession.

1. Introduction

Explanatory reductionism is the position which holds that, given a relatively higher-level phenomenon (or state, event, process, etc.), it can be reductively explained by a relatively lower-level feature (Kaiser 2015, 97; see also Sarkar 1998; Weber 2005; Rosenberg 2006; Waters 2008).[1] Though philosophers tend to have slightly different conceptions of the position, two central tenets of the position can still be extracted:[2]

[1] According to Sarkar (1998), explanatory reduction is an epistemological thesis which is distinguished from constitutive (ontological) and theory reductionism theses. Kaiser further distinguishes two sub-types of explanatory reduction: (a) "a relation between a higher-level explanation and a lower-level explanation of the same phenomenon" (2015, 97); (b) individual explanations, i.e., given a relatively higher-level phenomenon, it can be reductively explained by a relatively lowerlevel feature (*Ibid*., 97). This essay will focus on the second sub-type. Besides, when referring to levels I mean either hierarchical organization such as universities, faculties, departments etc., or functional organization such as organs, such as organs, tissues, cells etc. When referring to scales I mean spatial or temporal scaling where levels are not so clearly delimited.

[2] Similar summary of the position can be found in Sober (1999).

Single level preference: a phenomenon of interest can be fully explained by invoking features that reside at a single, well-defined level of analysis (e.g., molecular level in biology).

Lower-level obsession: lower-level features always provide the most significant and detailed explanation of the phenomenon in question, so a lower-level explanation is always better than a higher-level explanation.

Philosophers sometimes express these two tenets explicitly in their work. For example, Alex Rosenberg holds that "[…] there is a full and complete explanation of every biological fact, state, event, process, trend, or generalization, and that this explanation will cite only the interaction of macromolecules to provide this explanation" (Rosenberg 2006, 12). Marcel Weber expresses a similar idea in his explanatory hegemony thesis, according to which it's always some lower-level physicochemical laws (or principles) that ultimately do the explanatory work in experimental biology (Weber 2005, 18-50). John Bickle attempts to motivate a 'ruthless' reduction of psychological phenomena (e.g., memory) to the molecular level (Bickle 2003).

However, many philosophers have questioned the plausibility of the position on the basis of scientific practice (Hull 1972; Craver 2007; Bechtel 2010; Brigandt 2010; Hüttemann and Love 2011; Kaiser 2015). To counter that position, some authors have pointed to the relevance of an important practice that has not received sufficient attention before: multiscale or multilevel modeling or sometimes called integrative modeling approach,

where a set of distinct models ranging over multiple levels or scales—including the macro-phenomenon level/scale—are involved in explaining a (often complex) phenomenon of interest (Mitchell 2003, 2009; Craver 2007; Brigandt 2010, 2013a, 2013b; Knuuttila 2011; Batterman 2013; Green 2013; O' Malley et al. 2014; Green and Batterman 2017). Often these models work together by providing diverse constraints on the potential space of representation (Knuuttila and Loettgers 2010; Knuuttila 2011; Green 2013).

This multilevel modeling surely casts some doubt on explanatory reductionism, for it seems unclear what reductively explains what—all those facts in the set of models ranging over different levels/scales are involved in doing some explanatory work. However, there is a species of multilevel modeling that has slipped away from most philosophers' sights: mixed-effects modeling (MEM hereafter)—also called multilevel regression modeling, hierarchical linear modeling, etc.—in which a single model incorporating simultaneously two (or even more) levels of variables is used to explain a phenomenon. For a mixedeffects model to explain, features of the so-called reducing and reduced levels must be simultaneously incorporated into the model, that is, they must go hand in hand.

MEM deserves special attention because it sheds new light on the reductionism-antireductionism debate by showing that (a) a mixed-effects model violating the two central tenets of explanatory reductionism can provide successful explanation, and (b) a single mixed-effects model without integrating with other epistemic means can also provide such successful explanation. Therefore, MEM first further challenges the explanatory

reductionist position, and second offers a novel perspective bolstering the multilevel/multiscale integrative approach discussed by many philosophers.

The essay proceeds as follows. Section 2 discusses the challenges faced by the traditional single-level modeling approach, and examines the reasons why the MEM approach is preferable in dealing with these challenges. Section 3 describes a MEM practice using a concrete model. Section 4 elaborates on the implications of MEM for the explanatory reductionism debate. Finally, Section 5 considers potential objections to my viewpoint.

2. Challenges to Reductive Explanatory Strategies

In many fields (e.g., biological, social and behavioral sciences) scientists find that the data collected show an intrinsically hierarchical or nested feature. Consider a simple example: we might be interested in examining relationships between students' achievement at school (A hereafter) and the time they invest in studying (T).[3] In conducting such a research, we might collect data from different classes
(say 5 classes in total), with each class providing the same number of samples (say 10 students in each class). The data collected among classes might be taken for granted to be independent. Then we may use certain traditional statistical techniques such as ordinary least-squares (OLS) to analyze the data and build a linear relationship between A and T.

[3] For scientific studies of this kind, see Schagen (1990), Wang and Hsieh (2012), and Maxwell et al. (2017).

However, this single-level reductive analysis can lead to misleading results, because it ignores the possibility that students within a class may be more similar to each other in important aspects than students from different classes. In other words, each group (class) may have its own features relevant to the relationship between A and T that the other groups lack. Hence, the data collected from the students are in fact not independent, i.e., the subjects are not randomly sampled, because the individuals (students) are clustered within groups (classes). In technical terms, we say our analysis may fall prey to the *atomistic fallacy* where we base our analysis solely on the individual level—i.e., we reduce all the grouplevel features to the individuals. Therefore, traditional OLS techniques such as multiple regression cannot be employed in this context, because the case under consideration violates a fundamental assumption of these techniques: the independence of observations (Nezlek 2008, 843).

Conversely, we may face the same problem the other way around if we fail to consider the inherently nested nature of the data. Consider the studentachievement-at-school case again. We may observe that in classes where the time of study invested by students is very high, the achievements of the students are also very high. Given such an observation, we may reason that students who invest a lot of time in studying would be more likely to get higher achievements at school. However, this inference commits the *ecological fallacy*, because it attributes the relationship observed at the group-level to the individual-level (Freedman 1999). The individuals may exhibit within-group differences that the single group-level analysis fails to capture. In technical terms, this inference flaws because it reduces the variability in

achievement at the individual-level to a group-level variable, and the subsequent analysis is solely based on group's mean achievement results (Heck and Thomas 2015, 3). Again, traditional statistical techniques such as multiple regression cannot be employed in this context.

In sum, a single-level modeling approach that disrespects the multilevel data structure can commit either an atomistic or an ecological fallacy. Confronted with these problems, one response is to 'tailor' the traditional statistical techniques by, e.g., adding an effect variable to the model which indicates the grouping of the individuals. However, many have argued that this approach is unpromising because it may give rise to enormous new problems (Luke 2004; Nezlek 2008; Heck and Thomas 2015). Alternatively, scientists have developed a new framework that takes the multilevel data structure into full consideration, i.e., the MEM approach, to which we now turn.

3. Case Study: A Mixed-Effects Model

Depending on different conceptual and methodological roots we have two broad categories of MEM approaches: the multilevel regression approach and the structural equation modeling approach. The former usually focuses on direct effects of predictor variables on (typically) a single dependent variable, while the latter usually involves latent variables defined by observed indicators (for details see Heck and Thomas 2015). For the purpose of this essay's arguments, I will concentrate on the first kind.

Consider the student-achievement-at-school example again. Since students are typically clustered in different classes, a student's achievement at school may be both influenced by her own features (e.g., time invested in studying) and her class's features (e.g., size of the class). Hence here comes two levels of analysis: the individual-level (level-1) and the group-level (level-2), and individuals ($i = 1, 2, \dots, N$) are clustered in level-2 groups ($j = 1,2, \dots, n$).[4] Now suppose that students' achievements at school are represented as scores they get in the exam.

The effect of time invested in studying on scores can be described as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij} \qquad (1)$$

where Y_{ij} refers to the score of individual i in the jth group, β_{0j} is a level-1 intercept representing the mean of scores for the jth group, β_{1j} a level-1 slope (i.e., different effects of study time on scores) for the predictor variable X_{ij}, and the residual component (i.e., an error term) ε_{ij} the deviation of individual i's score from the level-2 mean in the jth group. Equation (1) looks like a multiple regression model; however, the subscript j reveals that there is a group-level incorporated in the model. It can also be seen from this equation that both the intercept β_{0j} and slope β_{1j} can vary across the level-2 units, that is, different groups can have different intercepts and slopes.

[4] Note that, for instructive purposes, our case involves only two levels; however, the MEM approach can in principle be extended to many more levels.

The most remarkable thing of MEM is that we treat both the intercept and slope at level-1 as dependent variables (i.e., outcomes) of level-2 predictor variables. So here we write the following equations expressing the relationships between the level-1 parameters and level-2 predictors:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \quad (2)$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j} \quad (3)$$

where β_{0j} refers to the level-1 intercept in level-2 unit j, γ_{00} denotes the mean value of the level-1 intercept, controlling for the level-2 predictor W_j, γ_{01} the slope for the level-2 variable W_j, and u_{0j} the error (i.e., the random variability) for unit j. Also, β_{1j} refers to the level-1 slope in level-2 unit j, γ_{10} the mean value of the level-1 slope controlling for the level-2 predictor W_j, γ_{11} the effect of the level-2 predictor W_j, and u_{1j} the error for unit j.

Equations (2) and (3) have specific meanings and purposes. They express how the level-1 parameters, i.e., intercept or slope, are functions of level-2 predictors and variability. They aim to explain variations in the randomly varying intercepts or slopes by adding one (or more) group-level predictor to the model. These expressions are based on the idea that the group-level characteristics such as group size may impact the strength of the within-group effect of study time on scores. This kind of effect is called a *cross-level*

interaction for it involves the impact of variables at one level of a data hierarchy on relationships at another level. We will discuss this in detail in the next section.

Now we combine equations (1), (2) and (3) by substituting the level-2 parts of the model into the level-1 equation. We finally obtain the following equation:

$$Y_{ij} = [\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}X_{ij}W_j] + [u_{1j}X_{ij} + u_{0j} + \varepsilon_{ij}] \qquad (4)$$

This equation can be simply understood that Y_{ij} is made up of two components: the fixed-effect part expressed by the first four terms and the random-effect part expressed by the last three terms. Note that the term $\gamma_{11}X_{ij}W_j$ denotes a crosslevel interaction between level-1 and level-2 variables, which is defined as the impact of a level-2 variable on the relationship between a level-1 predictor and the outcome Y_{ij}. We have 7 parameters to estimate in (4), they are four fixed effects: intercept, within-group predictor, between-group predictor and cross-level interaction, two random effects: the randomly varying intercept and slope, and a level-1 residual.

Now a mixed-effects model has been built, and the next step is to estimate the parameters of the model. However, we will skip this step and turn to explore the philosophical implications of the modeling practice relevant to the explanatory reductionism debate.

4. Implications for the Explanatory Reductionism Debate

Looking closely into the MEM practice, we find that a couple of important philosophical implications for the explanatory reductionism debate can be drawn.

4.1. All levels are indispensable

The first, and most obvious, feature of MEM is that it routinely involves many levels of analysis in a single model, and all these levels are indispensable to the model in the sense that no level can be reduced to or replaced by the other levels.

These levels consist of both the so-called reducing level in the reductionist's terminology, typically a lower-level that attempts to reduce another level, and the reduced level, typically a higher-level to be reduced by the reducing level. In our student-achievement-at-school case, for example, a reductionist may state that the group-level will be regarded as the reduced level whereas the student-level as the reducing level.

The indispensability of each level in the model can be understood in two related ways. First, due to the nested nature of data, only when we incorporate different levels of analyses to the model can we avoid either the atomistic or ecological fallacy discussed in Section 2. As discussed in the studentachievement-at-school example where students are clustered in different classes (in the manner that students from the same class may be more similar to each other in important aspects than students from different classes),

reducing all the analyses to the level of individual students can simply miss the important information associated with group-level features and thus lead to misleading results. Although it's true that the problem might be partially mitigated by tailoring traditional single-level analytical techniques such as multiple regression, it's also true that this somewhat ad hoc maneuver can simply bring about various new vexing and recalcitrant issues (Luke 2004; Nezlek 2008; Heck and Thomas 2015).

Second, the problem can also be viewed from the perspective of identifying explanatory variables. In building a mixed-effects model, the main consideration is often to find a couple of variables that may play the role of explaining the pattern or phenomenon observed in the data. Here a modeler must be clear about how to assign explanatory variables, for instance, she must consider if there are different levels of analyses and, if so, which explanatory variables should be assigned to what levels, and so on. These considerations may come before her model building because of background knowledge, which paves the way for her to develop a conceptual framework for investigating the problem of interest. However, without such a clear and rigorous consideration of identifying and assigning multilevel explanatory variables, an analysis can flaw simply because it confounds variables at different levels.

Respecting the multilevel nature of explanatory variables has another advantage: "Through examining the variation in outcomes that exists at different levels of the data hierarchy, we can develop more refined theories about how explanatory variables at each level contribute to variation in outcomes" (Heck and Thomas 2015, 33). In other words, in respecting the

multilevel nature of explanatory variables, we get a clear idea of how, and to what degrees, explanatory variables at different levels contribute to variation in outcomes. If these variables do contribute to variation in outcomes, as it always happens in MEM, then the situation suggests an image of *explanatory indispensability*: all the explanatory variables at different levels are indispensable to explaining the pattern or phenomenon of interest.

Given these considerations, therefore, one implication for the explanatory reductionism debate becomes clear: it isn't always the case that, given a relatively higher-level phenomenon it can be reductively explained by a relatively lowerlevel feature. Rather, in cases where the data show a nested structure or, put differently, the phenomenon suggests multilevel explanatory variables, we routinely combine the higher-level with the lower-level in a single (explanatory) model. As a result, one fundamental tenet of explanatory reductionism is violated:
single level preference.

4.2. Interactions between levels

Another crucial feature of multilevel modeling is its emphasis on *a cross-level interaction*, which is defined as

> "The potential effects variables at one level of a data hierarchy have on relationships at another level […]. Hence, the presence of a cross-level interaction implies that the magnitude of a relationship observed within

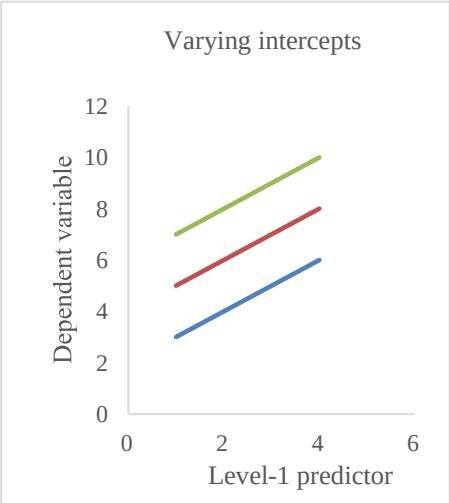
groups is dependent on contextual or organizational features defined by higher-level units". (Heck and Thomas 2015, 42-43)

Remember that there is a term $\gamma_{11}X_{ij}W_j$ in our mixed-effects model discussed in Section 3, which indicates the cross-level interaction between the group-level and the individual-level. More specifically, this term can be best construed as the impact of a group-level variable, e.g., group size, upon the individual-level relationship between a predictor, e.g., study time, and the outcome, e.g., students' scores.

The cross-level interaction points to the plain fact that an organization or a system can somehow influence its members or components by constraining how they behave within the organization or system. This doesn't necessarily imply top-down causation (Section 5.3 will turn back to this point). Within the context of scientific explanation, however, it does imply that it isn't simply that characteristics at different levels separately contribute to variation in outcomes, but rather that they interact in producing variation in outcomes. In other words, the pattern or phenomenon to be explained can be understood as generated by the interaction between explanatory variables at different levels. Therefore, to properly explain the phenomenon of interest, we need not only have a clear idea of how to assign explanatory variables to different levels but also an unequivocal conception of whether these explanatory variables may interact.

Different models can be built depending on different considerations of the cross-level interaction. To see this, consider the student-achievement-at-school example again. In some experiment setting we may assume that there was no

cross-level interaction between group-level characteristics and the individuallevel relationship (between study time and scores). In such a situation, we kept the effect of individual study time on scores the same across different classes, i.e., we kept the slope constant across classes. In the meanwhile, we treated another group-level variable (i.e., intercept) as varying across classes, i.e., different classes have different average scores. So, this is a case where we have a clear idea of how to assign explanatory variables but no consideration of the cross-level interaction. Nonetheless, in a different experiment setting we may assume that there existed cross-level interaction, and hence the effect of individual study time on scores can no longer be kept constant across different classes. At the same time, we treated another group-level variable (i.e., intercept) as varying across classes. Hence, this is a case where we have both a clear idea of how to assign explanatory variables and a consideration of the cross-level interaction. Corresponding to these two different scenarios, two different mixed-effects models can be built, as shown below:



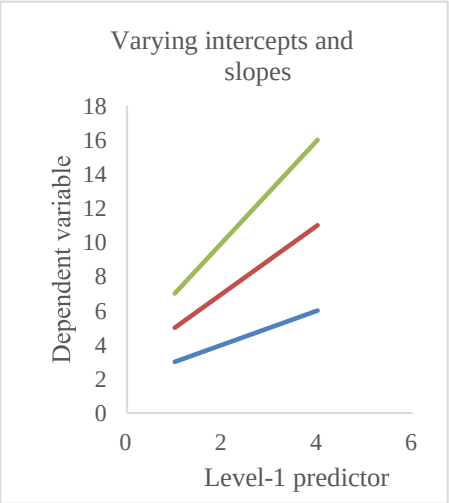


Figure 1. Two different models showing varying intercepts or varying intercepts and slopes, respectively. Three lines represent three classes. This figure is adapted from Luke (2004, 12).

Given such a cross-level interaction, therefore, the explanatory reductionist position has been further challenged. This is because any reductive explanation that privileges one level of analysis—usually the lower-level—over the others falls short of capturing this kind of interaction between levels. If they fail to do so, then they are missing important terms relevant to explaining the phenomenon of interest. As a consequence, a mixed-effects model involving interactions between levels simultaneously violates the two fundamental pillars of explanatory reductionism: first, it violates single level preference because it involves multilevel explanatory variables in explaining phenomena, and second, it violates lower-level obsession because it privileges no levels—all levels are interactively engaged in producing outcomes.

5. Potential Objections

This section considers two potential objections.

5.1. In-principle argument

One argument that resurfaces all the time in the reductionism-versusantireductionism debate is the in-principle argument, the core of which is that even if reductive explanations in a field of study are not available for

the time being, it doesn't follow that we won't obtain them someday (e.g., Sober 1999; Rosenberg 2006). Therefore, according to some reductionists, the gap between current-science and future-science is simply a matter of time, for advancement in techniques, experimentation and data collecting can surely fill in the gap.

However, I think the argument flaws. To begin with, advancement in techniques, experimentation and data collecting isn't always followed by reductive explanations. For example, in our MEM discussed in Section 3, even if the data about the individual-level is available and sufficiently detailed, it isn't the case that we explain the phenomenon of interest in terms of the data from the individual-level alone. Consider another example: in dealing with problems associated with complex systems in systems biology, even though large-scale experimentation (e.g., via computational simulation) can be conducted and high throughput data arranging over multiple scales/levels can be collected, a bottomup reductive approach must be integrated with a top-down perspective so as to produce useful explanations or predictions (Green 2013; Green and Batterman 2017; Gross and Green 2017).

Nevertheless, reductionists may reply that the situations presented above only constitute an in-practice impediment, for it doesn't undermine the *possibility* that lower-level reductive explanations, typically provided by some form of 'final science', will be available someday. Let us dwell on the notion of possibility a bit longer. The possibility here may be construed as a *logical possibility* (Green and Batterman 2017, 21; see also Batterman 2017). Nonetheless, if it's merely logically possible that there will be some final science providing only reductive explanations, then nothing can exclude

another logical possibility that there will be some ‘mixed-science’ providing only multilevel explanations. After all, how can we decide which logical possibility is more possible (or logically more possible)? I doubt that logic alone could provide anything useful in justifying which possibility is more possible, and that appealing to logical possibility could offer anything insightful in helping us understand how science proceeds. As Batterman puts, “Appeals to the possibility of *in principle* derivations rarely, if ever, come with even the slightest suggestion about how the derivations are supposed to go” (2017, 12; author’s emphasis).

Another interpretation of possibility may be associated with real possibilities, referring to the actual cases of reductive explanations happening in science. Unfortunately, I don’t think the real scenario in science speaks for the reductionist under this interpretation. Though it’s impossible to calculate the absolute cases of non-reductive explanations occurring in science, a cursive look at scientific practice can tell that a large portion of scientific explanations proceeds in a nonreductive fashion, as suggested by multilevel modeling (Batterman 2013; Green 2013; O’ Malley et al. 2014; Green and Batterman 2017; Mitchell and Gronenborn 2017). Moreover, even in areas such as physics which was regarded as a paradigm for the reductionist stance, progressive explanatory reduction doesn’t always happen (Green and Batterman 2017; Batterman 2017).

In sum, we have shown that the in-principle argument fails for it neither offers help in understanding how science proceeds if it’s construed as

implying a logical possibility, nor goes in tune with scientific practice if it's construed as implying real possibilities.

5.2. Top-down causation

In Section 3 we have shown that there is a cross-level interaction taking the form that higher-level features may impact lower-level features. A worry arises: Does this imply top-down causation?

My answer to this question is twofold. First, it's clear that this short essay isn't aimed to engage in the philosophical debate about whether, and in what sense, there exists top-down causation (see Craver and Bechtel 2007; Kaiser 2015; Bechtel 2017). Second, what we can do now is to show that the cross-level interaction is a clear and well-defined concept in multilevel modeling. It unambiguously means the constraints on the lower-level processes exerted by the higher-level parameters (Green and Batterman 2017). In our multilevel modeling discussed in Section 3, we have shown that group-level features may impact some individual-level features through the way that each group possesses its own feature relevant to explaining the differences at the individual-level across groups. This idea is incorporated into the mixed-effects model by assigning some explanatory variables to the group-level and a cross-level interaction term to the model.

The idea of cross-level-interaction-as-constraint is widely accepted in multilevel modeling broadly construed, where constraint is usually expressed in the form of initial and/or boundary conditions. For example, in modeling cardiac rhythms, due to "the influences of initial and boundary conditions on

the solutions of the differential equations used to represent the lower level process" (Noble 2012, 55; Cf. Green and Batterman 2017, 32), a model cannot simply narrowly focus on the level of proteins and DNA but must also consider the levels of cell and tissue working as constraints. The same story happens in cancer research, where scientists are advocating the idea that tumor development can be better understood if we consider the varying constraints exerted by tissue (Nelson and
Bissel 2006; Shawky and Davidson 2015; Cf. Green and Batterman 2017, 32).

6. conclusion

This essay has shown that no-reductive explanations involving many levels predominate in areas where the systems under consideration exhibit a hierarchical structure. These explanations violate the fundamental pillars of explanatory reductionism: single level preference and lower-level obsession. Traditional single-level reductive approaches fall short of capturing systems of this kind because they face the challenges of committing either the atomistic or ecological fallacy.

References

Batterman, Robert. 2013. The "Tyranny of Scales." In *The Oxford Handbook of Philosophy of Physics*, ed. Robert Batterman, 255-286. Oxford: Oxford University Press.

———. 2017. "Autonomy of Theories: An Explanatory Problem." *Noûs* 1-16.

Bechtel, William. 2010. "The Downs and Ups of Mechanistic Research: Circadian Rhythm Research as an Exemplar." *Erkenntnis* 73:313–328.

———. 2017. "Explicating Top-Down Causation Using Networks and Dynamics." *Philosophy of Science* 84:253–274.

Bickle, John. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.

Brigandt, Ingo. 2010. "Beyond Reductionism and Pluralism: Toward an Epistemology of Explanatory Integration in Biology." *Erkenntnis* 73 (3): 295311.

———. 2013a. "Explanation in Biology: Reduction, Pluralism, and Explanatory Aims." *Science and Education* 22:69–91.

———. 2013b. "Integration in Biology: Philosophical Perspectives on the Dynamics of Interdisciplinarity." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:461–465.

Craver, Carl. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, Carl, and William Bechtel. 2007. "Top-down Causation without TopDown Causes." *Biology and Philosophy* 22:547–563.

Freedman, David. 1999. “Ecological Inference and the Ecological Fallacy.” In *International Encyclopedia of the Social and Behavioral* Sciences, vol. 6, ed. Neil Smelser, and Paul Baltes, 4027–4030. New York: Elsevier.

Green, Sara. 2013. “When One Model Isn’t Enough: Combining Epistemic Tools in Systems Biology.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:170–180.

Green, Sara, and Robert Batterman. 2017. “Biology Meets Physics: Reductionism and Multi-Scale Modeling of Morphogenesis.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 61:20–34.

Gross, Fridolin, and Sara Green. 2017. “The Sum of the Parts: Large-Scale Modeling in Systems Biology.” *Philosophy, Theory, and Practice in Biology* 9: (10).

Heck, Ronald, and Scott Thomas. 2015. *An Introduction to Multilevel Modeling*

Techniques (3rd Edition). New York: Routledge.

Hull, David. 1972. “Reductionism in Genetics—Biology or Philosophy?” *Philosophy of Science* 39 (4): 491-499.

Hüttemann, Andreas, and Alan Love. 2011. “Aspects of Reductive Explanation in Biological Science: Intrinsicality, Fundamentality, and Temporality.” *British Journal for the Philosophy of Science* 62 (3): 519-549.

Kaiser, Marie. 2015. *Reductive Explanation in the Biological Sciences*. Springer.

Knuuttila, Tarja. 2011. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” *Studies in History and Philosophy of Science Part A* 42:262–271.

Luke, Douglas. 2004. *Multilevel Modeling*. London: SAGE Publications, Inc.

Maxwell, Sophie, Katherine Reynolds, Eunro Lee, et al. 2017. “The Impact of School Climate and School Identification on Academic Achievement: Multilevel Modeling with Student and Teacher Data.” *Frontiers in Psychology* 8:2069.

Mitchell, Sandra. 2003. *Biological Complexity and Integrative Pluralism.* Cambridge: Cambridge University Press.

———. 2009. *Unsimple Truths: Science, Complexity, and Policy*. Chicago: University of Chicago Press.

Nezlek, John. 2008. “An Introduction to Multilevel Modeling for Social and Personality Psychology.” *Social and Personality Psychology Compass* 2/2 (2008):842–860.

Noble, Daniel. 2012. “A Theory of Biological Relativity: No Privileged Level of Causation.” *Interface Focus* 2(1):55–64.

O’Malley Malley, Ingo Brigandt, Alan Love, et al. 2014. “Multilevel Research Strategies and Biological Systems.” *Philosophy of Science* 81:811–828.

Rosenberg, Alex. 2006. *Darwinian Reductionism, or How to Stop Worrying and*

Love Molecular Biology. Chicago: University of Chicago Press.

Sarkar, Sahotra. 1998. *Genetics and Reductionism*. Cambridge: Cambridge University Press.

Schagen, I. P. 1990. “Analysis of the Effects of School Variables Using Multilevel Models.” *Educational Studies* 16:61–73.

Shawky, Joseph, and Lance Davidson. 2015. “Tissue Mechanics and Adhesion during Embryo Development.” *Developmental Biology* 401(1):152–164.

Sober, Elliot. 1999. “The Multiple Realizability Argument against Reductionism.” *Philosophy of science* 66:542–564.

Wang, Yau-De, and Hui-Hsien Hsieh. 2012. “Toward a Better Understanding of the Link Between Ethical Climate and Job Satisfaction: A Multilevel Analysis.” *Journal of Business Ethics* 105:535–545.

Waters, C. Kenneth. 2008. “Beyond Theoretical Reduction and Layer-Cake Antireduction: How DNA Retooled Genetics and Transformed Biological Practice”. In *The Oxford Handbook of Philosophy of Biology*, ed. Michael Ruse, 238-262. New York: Oxford University Press.

Weber, Marcel. 2005. *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

Antireduction: How DNA Retooled Genetics and Transformed Biological Practice". In *The Oxford Handbook of Philosophy of Biology*, ed. Michael Ruse, 238-262. New York: Oxford University Press.

Weber, Marcel. 2005. *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

Two Ways of Mental State Attribution

Wuminyang

Abstract: There are different explanations for how we attribute minds to other persons or things. Based on the dual-process model of human cognition, Arico, Fiala, Goldberg, and Nichols put forward the so-called 'Agency Model'[1] to explain how we make inferences about others' mental states based on low-road cognitive processes. This paper aims to modify the Agency model by distinguishing two kinds of stimuli that trigger our mental attributions: the literal words and the concrete scenario. The Agency Model can be accepted when our mental states attributions operate without real scenarios. If the stimuli are just literal words and the words refer to agents, then we are inclined to attribute a wide range of mental states to the referents of the concepts. The Agency model is defective when the stimuli involve concrete scenarios, because in such cases we only attribute context-related mental states rather than mental states in general. Experiments were designed to test the new model. To conclude, a possible worry for the new model and its implications to experimental philosophy will be discussed.

Key words: Agency model; dual-process model; other minds

1. The problem of other minds and the Agency Model

The problem of other minds concentrates on answering how we can justify our belief that others have minds. New elucidations to the problem have arisen with the development of psychology. Experiments have been done to clarify what factors lead us to think others have minds, and how we attribute minds to nonhuman objects. All the discoveries are closely connected to the problem of other minds and our understanding of attributions of consciousness. There are different ways to categorize different answers to the problem. One way to manifest the differences is by appealing to the dual-process model of cognition[2]. This model divides our cognition into two systems:

The operations of System 1 are typically fast, automatic, effortless, associative, implicit (not available to introspection), and often emotionally charged; they are also governed by habit and are therefore difficult to control or modify. The operations of System 2 are slower, serial, effortful, more likely to be consciously monitored and deliberately controlled (Kahneman, 2003, p. 698).

Based on the distinction, we can also divide the answers to the problem of other minds into two groups. Analogy theory and IBE (Inference to the Best Explanation) theory seem to be based

[1] Arico, A., Fiala, B., Goldberg, R., & Nichols, S. (2011). The folk psychology of consciousness. Mind & Language, 327-353.

[2] Kahneman, D. 2003: A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, 58, 698.

on system 2. Both theories require deliberations when such inferences are made: we chew over whether others have minds by comparing them with our own experience, or weigh carefully whether our attributions of minds to others gives the best explanation for others' behaviors. While these two philosophical answers are clearly based on system 2, a number of empirical experiments pay attention to our direct and intuitive inclination to attribute mental states to nonhuman objects, which can safely be understood as the product of system 1. One account of this kind is the Agency model advanced by Arico, Fiala, Goldberg, and Nichols. They focus on 3 main cues which contribute to the triggering of mental attributions: the presence of eyes, motion trajectories, and contingent interaction. Once one or some of these cues are presented to us in an object, our concept of an 'agent' will be triggered. Once we treat the object as an agent, we will have the inclination to attribute *a wide range of* mental states to the object. Figure 1 below is cited from their paper to help us understand the model.

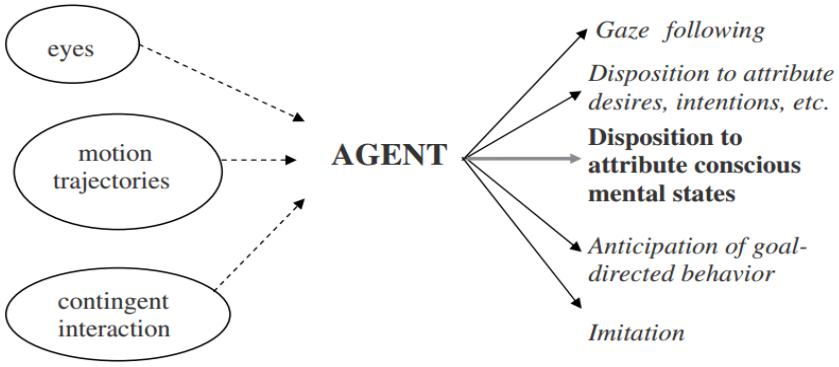


Figure 1: Network of AGENT concept

They argue that:

The Agency Model claims that triggering the low-road process and categorizing an entity as an AGENT disposes us to attribute both intentional and phenomenal mental states, as well as both valenced and unvalenced mental states. That is to say, low-road processing automatically facilitates ***a wide range of*** *mental state attributions.*[3]

Debates mainly focus on whether or not we are inclined to attribute 'a wide range of' mental states to the object if the object has one or some of these three traits. Justin Sytsma raised a potential worry for the model based on his experiment with Edouard Machery . They found that participants tended to attribute 'seeing red' but not 'feeling pain' to Jimmy the robot, even though Jimmy has all these three traits[4]. According to the Agency Model, the robot should be classified as an agent and people should be disposed to attribute a wide range of mental states to Jimmy including both 'seeing red' and 'feeling pain'. Although Sytsma later convincingly

[3] Fiala , B.,Arico, A. and Nichols,S.(2014).You, Robot.*Current Controversies in Experimental Philosophy,*36.

[4] Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies*, 151, 299–327.

argued that the Agency Model isn't in conflict with their results[5], the experiments still cast doubt on the Agency Model: People seem to attribute some, but not most possible mental states, to the robot despite the robot's having all the relevant cues.

We can see from figure 1 that the cues on the left seem to be real traits or patterns when the object is presented to us. However, Arico, Fiala, Goldberg, and Nichols used abstract words like 'bird' or 'insect' as the stimuli. Can such stimuli really arouse our impressions of the three cues so that we treat the reference of the abstract concept as an agent and then attribute mental states to it? Based on what we've discussed above, some more specific questions can be raised here: under what conditions can the agency model be used to explain our inclination to attribute mental states to agents? Is there a difference between attributions of consciousness without real objects presented and under real circumstances? If this distinction exists, maybe it's more natural to suppose we are more inclined to attribute context-related mental states, rather than a wide range of mental states to the object under concrete circumstances.

2.The modified agency model

The plan to modify the model is based on the division of two kinds of stimuli: one kind of stimuli involves real objects presented in concrete scenarios while the other only involves abstract words or sentences. To reflect this division, the new model should contains two levels, the abstract level and the concrete level: if the stimuli are words or sentences, the Agency model's claim can still be true that we are inclined to attribute most possible mental states to something as an agent; if the stimuli are concrete scenarios, the Agency model is defective and should be weakened so that we attribute only context-related mental states to the object. To summarize:

Our mental state attributions can be divided into two kinds in terms of the conditions we are under. Under one condition, if there is no real object when we we make mental states attributions, we make such inferences based on whether this object can be conceptually categorized as an agent. Once we treat the concept as the subclass of the agent, we attribute a wide range of mental states to the reference of the concept. Under the other condition, if we make mental state attributions in concrete scenarios, the kinds of mental states we attribute will depend on what traits or patterns we can recognize. Specific behavioral patterns are linked with specific mental states.

Figure 2 below can help to illustrate the idea. For convenience, we may name the left part as **the abstract level**, and the right as **the concrete level** of the new model. We may name the new model as the MA model (modified agency model) for convenience in what follows.

[5] Sytsma, J.(2013)The Robots of the Dawn of Experimental Philosophy of Mind. *Current Controversies in Experimental Philosophy,48-64.*

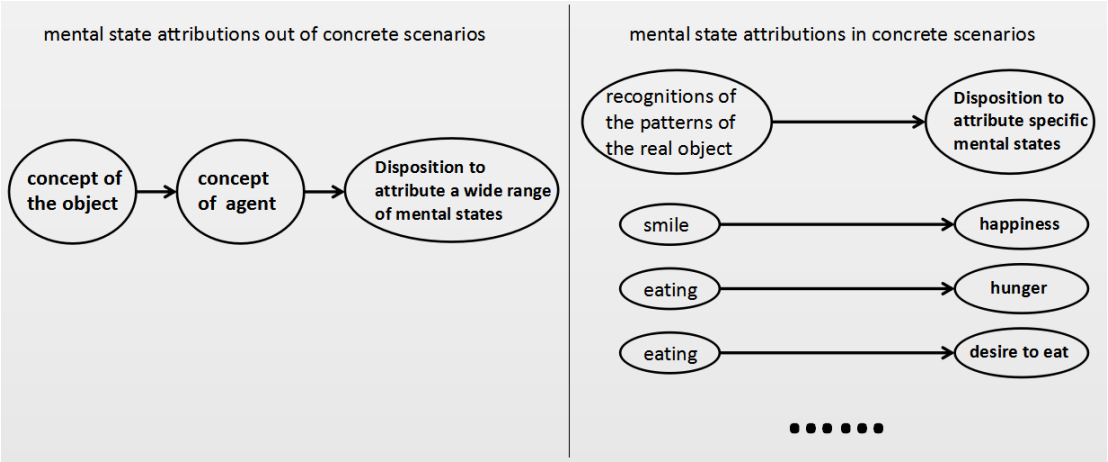


Figure 2: the modified agency model

We can make a couple **predictions** based on the MA model:

***First**, If the abstract level of the model is true, people will be inclined to attribute a wide range of mental states to agents and aren't inclined to attribute mental states to non-agents. Also, it's easier for people to attribute mental states compared with not attributing mental states when the agent concept is triggered, and it's easier not to attribute mental states compared with attributing mental states when the agent concept is not triggered.*

***Second**, If the concrete level of the model is true, people would be inclined to attribute context-related mental states rather than context-unrelated mental states to agents in concrete scenarios, and have inclinations to attribute few or no mental states to non-agents, because non-agents usually display no significant traits or behavioral patterns even if they are presented as real objects.*

*These two predictions aim to test the abstract level and the concrete level of the new model respectively. **In addition to these predictions**, we still need to test whether the division between these two levels are statistically significant. If the differences really exist, we can predict that, when the same object is presented in the form of an abstract concept and a real thing separately, people would attribute fewer mental states to the real thing than to the referent of the concept if the object is an agent; if the object is not agent at all, people are not inclined to attribute mental states to the object regardless of whether the object appears as an abstract concept or a real thing.*

3. Experiments

Goal: The experiment used a within-subjects design. The whole experiment is composed of two sub-experiments. Experiment 1 aims to test the first prediction mentioned above and Experiment

2 aims to test the second prediction. Comparing the results of these two experiments, we can test the third prediction that the division between these two levels is significant. To avoid the practice effect or other forms of interference that participants in Experiment 1 may bring to Experiment 2, the Experiment 2 was conducted at least 10 days after the same participant finished the Experiment 1.

Participants: Participants are 31 college students (19 male, 12 female, mean age = 23.5) who have not been exposed to our current issues from Renmin University and Capital Normal University. Only one of the participants is a Christian and all the rest have no religious beliefs. All participants were paid at least 5 yuan.

Design and materials: Both of these two experiments employed the reaction time paradigm. In both experiments, sentences are presented on the computer screen for participants to make judgements on whether the meaning of the sentences are appropriate to them according to their intuitions. The sentences are all statements saying that certain object has certain mental state, such as "甲虫感到疼痛" (beetle feels pain). Participants should give a 'Yes' or 'No' response as quickly as possible by pressing 'q' (for yes) or 'p' (for no) on the keyboard. However, in the first experiment, there are only such sentences for participants to make yes-or-no choices within 4 seconds, while in the second experiment, participants were asked to **watch a short video** before they made the choices. The sentences used in Experiment 1 and Experiment 2 are the same. There are 40 sentences in total for describing 5 object and 4 kinds of mental states under 2 conditions (40 = 5×4×2). The objects include mouse, beetle, mushroom, cloud and finally, a robot. Mental states are sensation (some are perception), emotion, desire and belief. For each match of an object and a type of mental state, there are two sentences, one is relevant and the other is irrelevant to the video.

In the first experiment, 32 sentences except those about the robot are presented to participants randomly. After making 32 yes-or-no choices, participants are asked to read a paragraph describing the behaviors of a robot. After reading the paragraph, the remaining 8 sentences concerning the robot are presented to them for yes-or-no choices randomly. The English version of the paragraph is attached in the appendix.

In the second experiment, 5 videos are added. After seeing the video, participants are presented with 8 sentences concerning the object in the video, which forms a group. To impel participants to make judgements according to their understanding of the video, a picture cut from the video showing the object is presented above the sentences. Also, to avoid the concepts of objects in the sentences inducing people to think on an abstract level, all concepts of objects are replaced by 'Ta' or 'Ta men', which are the Chinese pronunciations of all personal pronouns such as he, she, it, they. As mentioned above, half of the sentences are relevant to the content of the video while the other half are irrelevant. The content of the robot video has been

summarized in the paragraph mention above(refer to appendix).

All stimuli in the two experiments were developed on E-prime. The following chart shows us the flow of stimuli of each experiment. See Figure 3.

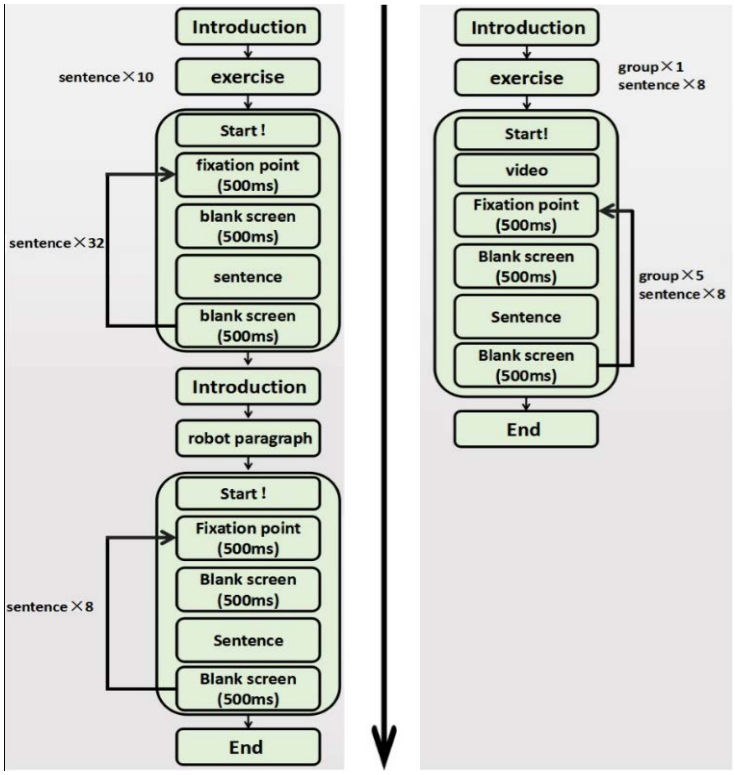


Figure 3: Flow chart of stimuli in Experiment 1 (left) and Experiment 2 (right)

Predictions: Based on our model, we can set standard answers for each experiment and calculate the grades of each participant to see to what extent their answers accord with our model. The higher grades participants get, the better our model reveals the fact. For Experiment 1, the grade of each participant is calculated by the following formula:

Grade 1 =[N(agent,Yes)+N(Non-agent,No)]/32*100

In the formula, N(agent,Yes) refers to the number of 'yes' responses the participant gives when the sentence concerns an agent. The rest can be understood in the same manner. Grade 1 doesn't include trials concerning the robot. Data of the robot in Experiment 1 is analyzed independently. For the second experiment, the grade of each participant is calculated by the following formula:

Grade 2 =[N(agent, relevant, Yes)+N(agent, irrelevant, No)+N(Non-agent, relevant, No)+N(Non-agent, irrelevant, No)]/40*100

In the formula, N(agent, relevant, Yes) refers to the number of 'yes' responses the participant gives when the sentence concerns an agent and the mental state under consideration is relevant

to the content of the video. The rest can be understood in the same manner. Notice that for non-agents, we included the number of participants' 'no' responses into the formula regardless of whether the sentence was relevant to the video, because, as mentioned above, non-agents seem to display no significant traits or behavioral patterns even if they are presented as real objects. More predictions have been shown in the end of Part Two.

Results:

For Experiment 1:

The average numbers and reaction times of yes-or-no choices are shown in the bar graphs below (Figure 4 and 5).

①Avg N(agent,Yes) = 5.15 > 4[6], which implies that participants are more inclined to attribute mental states to agents;

②Avg N(Non-agent, No) = 5.5 > 4,which implies that participants are more inclined not to attribute mental states to non-agents;

③Avg Grade 1 = 66.43 >50 (SD=17.75, t(30)=5.15, p<.001), which implies that the abstract level of the modified model can be accepted;

④Avg RT(Beetle, yes) [7]= 1331.82 < Avg RT(Beetle, No) = 1524.87 and Avg RT(Mouse, yes) = 1314.14 < Avg RT(Mouse, No) = 1582.30. This implies that it's easier for participants to attribute mental states to agents than reject it.

⑤Avg RT(Mushroom, yes) = 1428.30 < Avg RT(Mushroom, No) = 1445.86 and Avg RT(Cloud, yes) = 1279.85 < Avg RT(Cloud, No) = 1349.84. This result is in conflict with our prediction ⑤.

[6] If our concept of agent has no influence on our mental attribution inclinations, the average N(agent,Yes) is near 4 if the sample size is large enough, which supposes that the probability that people attribute mental states to agents equals to the probability that people reject mental state attributions to agents. Other predictions below are based on the same kind of hypotheses.

[7] RT(agent, yes) refers to the reaction time for participants to give a yes answer to the sentence concerning the robot. The rest can be understood in the same manner.

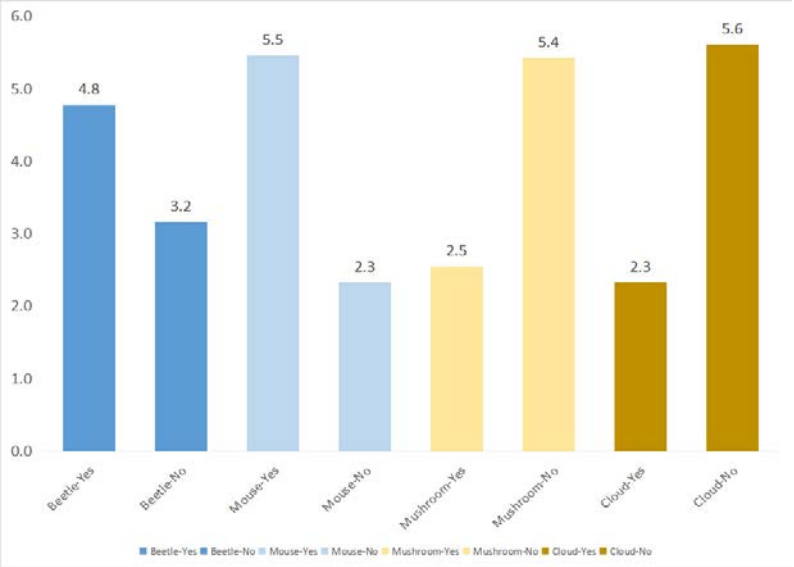


Figure 4: Average numbers of yes-or-no choices for each object

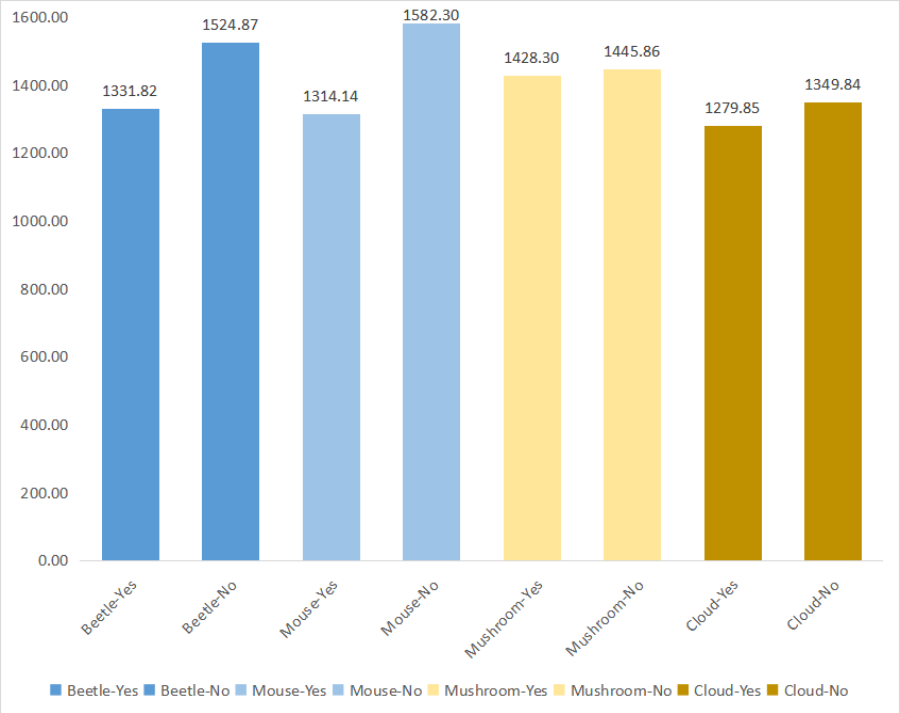


Figure 5: Average reaction times (ms) of yes-or-no choices for each object

For Experiment 2:

The average numbers of yes-or-no choices in terms of whether the sentences are relevant (r.) or irrelevant (irr.) to the video are shown in the bar graph below (See Figure 6).

①Avg N(mouse, relevant, Yes) = 3.35 >Avg N(mouse, irrelevant, Yes) =0.58; Avg N(robot,

relevant, Yes) = 2.71 >Avg N(robot, irrelevant, Yes) =0.10; Avg N(beetle, relevant, Yes) = 2.68 >Avg N(beetle, irrelevant, Yes) =0.55; So, in all trials about agents, participants are more inclined to attribute context-related than context-unrelated mental states to agents.

②N(mushroom, relevant, Yes) =2.52, N(mushroom, irrelevant, Yes) =0.48; N(cloud, relevant, Yes) =1.03, N(cloud, irrelevant, Yes) =0.55. This implies that participants are not inclined to attribute mental states to non-agents, regardless of whether the sentences are relevant to the video.

We may find some more interesting phenomena here. For both the mushroom and the cloud, N(mushroom, relevant, Yes)>N(mushroom, irrelevant, Yes) (SD=1.22, t(30)=9.24, p<.001) and N(cloud, relevant, Yes)>N(cloud, irrelevant, Yes) (SD=0.72, t(30)=3.72, p<.001), which implies that although participants are not inclined to attribute mental states to non-agents, *concrete scenarios really enhance the possibility of participants to attribute the context-related mental states to non-agents*.

③Avg Grade 2 =77.26 > 50 (SD=9.20, t(30)=16.49, p<.001), which implies that the concrete level of the modified model can be accepted, and the prediction here is even more statistically significant than that of Grade 1 in Experiment 1;

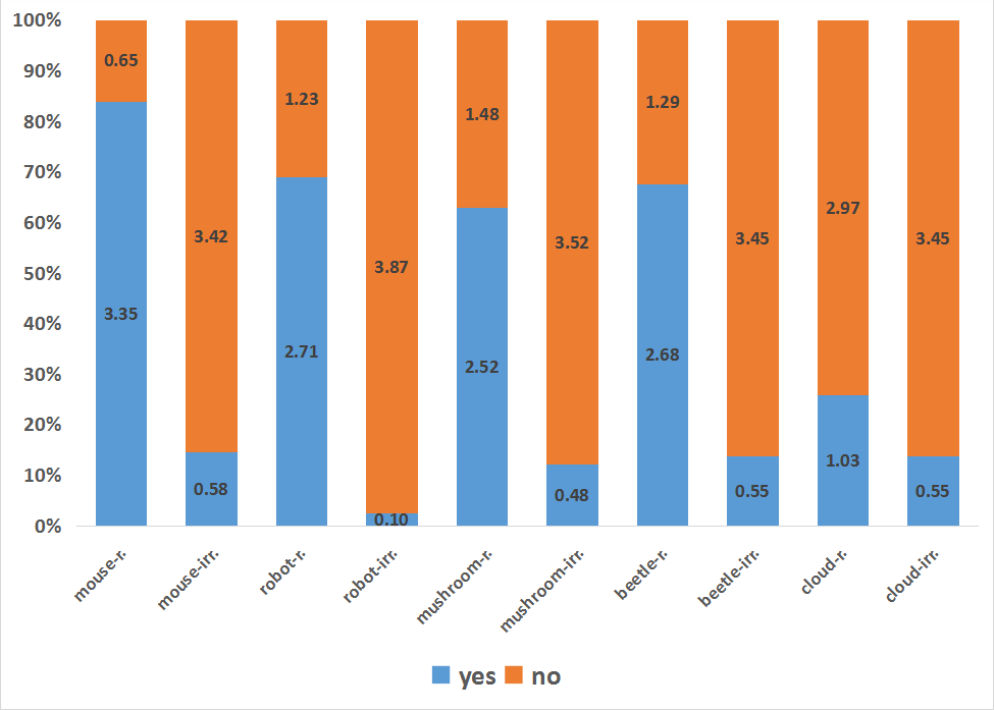


Figure 6: Average number of yes-or-no choices concerning the relevancy to the video

For comparing Experiment 1 and Experiment 2:

We still need to check whether the division between these two levels are statistically significant.

①Avg N(mouse, Yes) $_{exp.\ 1}$ = 5.5 > Avg N(mouse, Yes) $_{exp.\ 2}$ = 3.94 (SD=2.19, t(30)=3.86, p<.001); Avg N(beetle, Yes) $_{exp.\ 1}$ = 4.8 > Avg N(beetle, Yes) $_{exp.\ 2}$ = 3.23 (SD=2.28, t(30)=3.78, p<.001). This justifies the prediction that people attribute fewer mental states to the real thing than to the reference of the concept if the object is an agent, since they just attribute context-related mental states to the agent.

②Avg N(mushroom, Yes) $_{exp.\ 1}$ = 2.55 < Avg N(mushroom, Yes) $_{exp.\ 2}$ = 3.00 (SD=2.64, t(30)=-0.95, p=.175) while Avg N(cloud, Yes) $_{exp.\ 1}$ = 2.32 > Avg N(cloud, Yes) $_{exp.\ 2}$ = 1.58 (SD=2.18, t(30)=1.90, p=.034). This result cannot support prediction ② mentioned above.

For robot trials in Experiment 1 and Experiment 2:

①Avg N(robot, relevant, Yes)$_{exp.\ 1}$ = 2.39 < 4 and Avg N(robot, relevant, Yes)$_{exp.\ 2}$ = 2.71 < 4. which implies that people are not inclined to attribute context-unrelated mental states to the robot.

②Also, we can see that Avg N(robot, relevant, Yes)$_{exp.\ 2}$ > Avg N(robot, relevant, Yes)$_{exp.\ 1}$ (SD=0.95, t(30)=1.90, p<.05), which implies that real scenarios really tempt participants to attribute more mental states to the robot.

Conclusions: Most predictions are true according to the result, only the last point in **Prediction 1** and the second point of **Prediction 3** cannot be accepted. Since these two points are not essential to our whole model, we may cautiously accept the modified agency model.

4. Further discussions

4.1 Understanding the Robot Jimmy again

Since the debate attracting our attention here was originally provoked by applying the Agency model to predict how people attribute mental states to the robot 'Jimmy', we may consider if the robot is somehow different from other kinds of things. We should notice that most 'other things' we've talked about are all things whose concepts and exterior traits or patterns are consistent with each other. Animals or insects are such natural kinds that their concepts promise a series of exterior traits or patterns: being living things, being able to act and can sense the environment……, but we don't attach animate traits or patterns to water or stone. In other words, most agents in nature are living things and can exemplify different animate traits or patterns, while most non-agents in nature are inorganic substances from which we never expect such traits or patterns.

Robots are different from them. All robots are inorganic substances which conceptually implies that they cannot exemplify any animate traits or patterns. However, robots are designed

to exemplify such animate traits or patterns which is contradictory to its conceptual category in our minds. Furthermore, we have stereotyped images of natural things, we know what kinds of characteristics they may have based on our experience. But we have no such stereotyped images of robots until we see a real robot to judge what exterior traits it can have. This point is essential in terms of our MA model and has been supported by the result of our experiment. In brief, robots are inorganic substances so we are inclined to attribute no mental states to them based on abstract concepts. But once we are faced with a real robot and witness its abilities, we are still inclined to attribute some mental states to it according to the concrete scenario. Since the animate traits or patterns we recognize in robots are designed by humans, such traits or patterns may not be strong enough to push us into attributing the relevant mental states or to overcome our assumptions that robots are lifeless stuff.

This point may help to explain why participants in Sytsma and Machery's experiment are more inclined to attribute 'seeing red' than 'feeling pain' to the robot. Because what mental states are for and to what extent they are attributed to the robot quite depends on how strong the implication of the behavioral patterns is. We can imagine, if the robot Jimmy looks just like a real human, people can hardly reject their intuitions to attribute the relevant conscious mental states to it. So why are people more inclined to attribute 'seeing red' than 'feeling pain' to the robot? Perhaps, because they don't hear the screech of the robot or see the robot bleeding.

4.2 A worry for the new model

There is a widely accepted consensus that all experiments exploring our psychological or mental phenomena have to face the problem of external validity. External validity refers to the extent to which the results of a study can be generalized to other situations and to other people[8]. in terms of this problem, is there such a possibility that the concrete level of our new model is just the result of applying the mechanisms operating at the abstract level to the real world? If this is the case, what we do here might be quite trivial. That means there is only one mechanism processing all the stimuli which had been depicted by the original agency model. The concrete level of our model is nothing more than the application of the original agency model to the real world. So there are two possibilities: first, there are two different mechanisms that respectively process the semantic information and the episodic information; Second, there is only one mechanism, if we activate it conceptually, we get our abstract level of our new model; if we activate it under concrete circumstances, we get our concrete level of our new model. The following picture illustrates these two cases.

[8] The description is cite from Wikipedia. For more details, see https://en.wikipedia.org/wiki/External_validity.

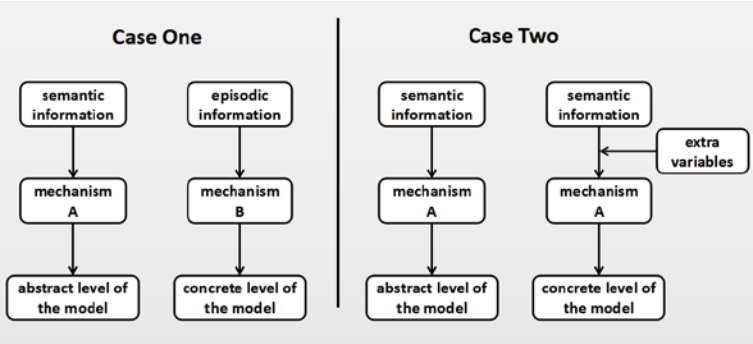


Figure 7: Two ways to understand the model

This worry is reasonable at first sight. Because it's very natural to ask if there are two special neural mechanisms underlying these two processes. But as mentioned above, these two processes can easily affect each other, so the neural mechanism in charge can hardly be distinguished completely. If we understand the 'mechanism' as 'neural mechanism', it's too obvious that these two processes have to share some common neural mechanisms in our brain. So the mechanism here may be better understood as a functional mechanism. If the two processes have different functions, we may divide them into two different mechanisms. Similar divisions can be seen in psychology. For example, our memory can be divided into episodic memory and semantic memory, but scientists believe that both are stored in the medial temporal lobes and hippocampal formation. Being located in the same parts of the brain doesn't prevent us from distinguishing them in terms of their different functions. Our division of the two levels of the model may be seen in the same way. Since there really exist significant differences between the input information and output behavior as shown in the experiment, we can still hold that our modification of the original model can be accepted.

4.3 Implications in experimental philosophy

To summarize our work, we try to show how we make different conscious state attributions based on the division of two kinds of stimuli. We are not alone. A similar approach has been adopted by some experimental philosophers to explore other important issues. For example, concerning our concepts about causation, Danks, Rose, and Machery have designed experiments to show that the way we understand causation depends on how the information is presented. Information presented by text or experience can bring out different judgments in people.[9] Regarding free will and responsibility, similar differences can be found between abstract stimuli and more concrete stimuli. [10]

[9] Danks, David, David Rose, and Edouard Machery. 2014. Demoralizing Causation. *Philosophical Studies* 171: 251–277.

[10] Nichols, S. & Knobe, J., "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions", Nouns, 41(4), 2007, pp.663-685.

This research reminds us that what we find under the experimental condition is quite different from how we make judgments in the real world. What we think when we answer questionnaires might be distant from how we act in our life. Noticing the division of abstract stimuli and concrete stimuli in experiments may help us to get closer to our 'true' intuitions, if our 'true' intuitions lie in our daily life.

Appendix

1. The paragraph of robot used in Experiment 1:

Cozmo is an intelligent robot toy developed by an American company. It has a screen, which can present different kinds of expressions; a video camera is hidden behind the screen for the robot to detect the external environment. It has a pair of apron wheel for moving and is equipped with a simple mechanical arm to touch and move surrounding things. Please imagine the following scenario: After the robot Cozmo is arranged in one corner of the room, it wears an expression as looking around. Then, its expression changes into a snigger and it moves to a sleeping dog in the room. It stops beside the tail of the dog, aims its screen to the tail and uses its mechanical arm to hit the tail at full tilt. The dog wakes up suddenly. Then the robot wears an expression of gloat on its screen.

Subjective Beliefs in Outcome Probability and Moral Decision in Moral Dilemmas

Song Fei

Abstract

Previous studies have found that the proportions of people who endorse utilitarianism decisions varied across different variants of the trolley dilemma. In this paper, we explored whether moral choices were associated with participants' beliefs of outcome probabilities in different moral dilemmas. Results of two experiments showed that participants' perceptions of outcome probabilities were significantly different between two dilemmas that were similar to the classical *switch* case and *footbridge* case. Participants' judgments of the outcome probabilities were significantly associated with their moral choices. The results also suggested that participants might not accept task instructions and thus did not perceive the outcomes in the dilemmas as certain. We argued that researchers who endorse descriptive tasks in moral reasoning research should be cautious about the findings and should take participants' beliefs in the outcomes into account.

Keywords**:** moral decision; probability judgment; moral reasoning; moral dilemma

Moral reasoning has been under long-term intellectual scrutiny. Recent psychological investigations of moral reasoning frequently employ moral dilemmas (Crockett, 2013; Cushman, Young, & Hauser, 2006; Greene et al., 2009; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Haidt, 2007; Moll & de Oliveira-Souza, 2007). The well-known trolley dilemma requires people to choose between two options: killing one person to save five or letting the five people die. Moral dilemmas such as this commonly engender conflict between two major approaches to moral reasoning: consequentialist and deontological approaches[1]. The consequentialist approach primarily concerns the outcome of each option and aims at choosing the one with the best outcome. By contrast, the deontological approach is concerned with whether an act is consistent with a moral principle or duty. In most studies, the choice of killing (directly or indirectly) one person in the trolley dilemma is taken as a result of maximizing the outcome utility (i.e., the number of people who will not die) and is often associated with the consequentialist approach. Not killing, on the other hand, is taken as a product of the deontological approach under which the action of killing is regarded as a deontological violation (Cummins & Cummins, 2012).

Previous studies revealed that people's moral judgments (i.e., the choice of which option is more morally acceptable) varied across different dilemmas (Cummins & Cummins, 2012; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). In the *switch* case, participants have to decide between flipping a switch to shift a trolley to a side-track on which one person is trapped, and doing nothing. The majority of participants perceived the choice of switching the trolley (consequently, killing the minority) as morally preferable. However, most participants preferred doing nothing (or not killing the minority) in other variants of the trolley dilemma, such as the well-known footbridge dilemma (e.g., Crockett, 2013; Greene et al., 2001, 2009; Lerner, Li, Valdesolo, & Kassam, 2014; Shenhav & Greene, 2014; Valdesolo & DeSteno,

[1] In the current paper, we use 'approach' and 'reasoning' interchangeably.

[2] Some may argue that, strictly speaking, the negative outcome should be that all six

2006; Youssef et al., 2012). In the footbridge dilemma, one is required to decide whether or not to push a fat man over a footbridge to stop a runaway trolley that would otherwise kill five people. It has been of great interest of researchers as to why there is a discrepancy in moral judgments between these two types of dilemmas.

A lot of studies have attributed the differences to certain features of moral dilemmas. For example, personal dilemmas (e.g., the footbridge case) involve harm with personal forces and thereby a stronger negative affect would be elicited, which has often been characterized as the main reason underlying the commonality of the preference for not killing the individual (Greene, Nystrom, Engell, Darley, & Cohen, 2004). A methodological problem with most studies is that they equate the choice of killing/not killing with the endorsement of the consequentialist/deontological reasoning approach, upon the presumption that participants perceived the outcomes in moral dilemmas as certain (Kortenkamp & Moore, 2014). However, participants may hold or be influenced by subjective beliefs regarding how likely it is an outcome would occur when a particular choice is taken. This consequently influences their expected utilities of the given choices in different moral dilemmas. The perceived uncertainty in the outcomes could reduce the conflict between deontological versus consequentialist approaches was reduced as the action of killing is no longer producing the best outcomes.

In this paper, we explored whether participants' choice preferences can be associated with subjective beliefs of the probabilities of the outcomes in the dilemmas. Choosing killing (hereafter *K*) or not killing (hereafter ~*K*) as the morally preferred choice does not necessarily reflect deontological or consequentialist approaches. Though ~*K* is usually taken as a result of deontological reasoning, a consequentialist may also choose ~*K* when the expected value of killing is lower than that of ~*K*. For instance, in the footbridge dilemma, one may assign a probability lower than 100% to the outcome that sees the trolley being stopped by the fat man. In this case, the aversion to "doing harm" in the footbridge dilemma could be taken as a result of deontological reasoning or as a consequence of the aversion to uncertainty (Rogers, Viding, & Chamorro-Premuzic, 2013).

Study

Our study examined how participants perceived the outcome possibilities in different types of dilemmas and how their judgments of outcome probabilities were associated with their moral choices. We utilized two dilemmas. One involved personal force while the other did not (see materials for details). In accordance with dilemmas used in previous studies, the dilemmas in the current study specified the positive outcome (or benefit) and the negative outcome for each of the two choices (*K* or ~*K*). The positive outcome of *K* was the survival of five people, while its negative outcome was the death of one person. On the other hand, the positive outcome of ~*K* was the survival of one person, while its negative outcome was the death of five people[2]. In

[2] Some may argue that, strictly speaking, the negative outcome should be that all six people die, and the positive outcome should be all six people survive. Because the probability that the five people die/survive differs from the probability that one person

each dilemma, participants provided probability judgments for the positive and negative outcomes from a given choice (one for each of the four parts of a dilemma, see below for more details).
We hypothesized that:

H1: Participants' judgments of the outcome probabilities would be significantly different between different dilemmas.
H2: Participants' moral choices would be significantly associated with their perceived outcome probabilities. Higher probabilities of positive outcomes and lower probabilities of negative outcomes given a particular choice would be associated with a higher likelihood of endorsing that choice.

Method

Participants and procedure

A total of 112 participants (85 females) were recruited via online crowd-sourcing service, CrowdFlower. Participants were aged between 19 and 71, with a mean age of 40.9 years (*SD* = 11.67). Participants were randomly assigned to one of two moral dilemma scenarios (described below). They read the consent information and completed the demographical questionnaire and the moral judgment task in order.

Materials

One of the two vignettes described a *car* dilemma, where the participant was asked to imagine they were driving a truck approaching a sharp turn near a cliff. A car of five passengers suddenly stopped in front of the truck. The participant had two options, of which the *K* choice was to turn the truck into one bystander and the other option was to let the truck hit the car with the five people inside. The other vignette described a *hostage* dilemma, in which the subject was passing by a cliff with another innocent person. The participant was threatened by a gangster who had captured five hostages. The participant had two options: to push the other person over the cliff or to let the gangster shoot the five hostages[3]. The car and hostage dilemmas differed in terms of whether the agent (the participant) had physical contact with the victim. This personal/impersonal distinction was similar to one of the differences between the trolley and the footbridge dilemmas.

The description of the two dilemmas did not contain any probabilistic information, but indicated that the individual victim (passerby or the innocent person) would die if *K* was chosen, while the five victims (five passengers or hostages) would die if ~*K* was chosen. After reading the assigned dilemma, participants were asked, "Which action do you think is morally better?" We used this question to focus participants on the issue of morality. Baron (2013) argued that conventional moral judgment questions that use terms such as "permitted" or "wrong" may draw participants' attentions to law or convention and away from morality. In addition, we avoided

---

dies/survives for each action taken, it does make sense to ask participants to estimate whether "all six people die/survive".

[3] The details of materials as well as example illustrations are available in online supplemental materials at http://goo.gl/hknhMJ

asking questions regarding which action the participant would choose to distinguish moral judgment from preference.

After making the moral choice, participants provided probability judgments for four possible outcomes. As mentioned earlier, the positive outcome (PO) given *K* is the survival of five people, while its negative outcome (NO) is the death of the individual. The positive outcome given ~*K* is the survival of the individual, while its negative outcome is the death of the five people. Participants provided judgments for the following probabilities:

1) *P*(PO|*K*): the probability that five people would survive given *K* is chosen;
2) *P*(NO|*K*): the probability that one person would die given *K* is chosen;
3) *P*(NO|~*K*): the probability that five people would die given ~*K* is chosen;
4) *P*(PO|~*K*): the probability that one person would survive given ~*K* is chosen.

Results

For the moral choice, participants in the car dilemma were significantly more likely to choose *K* than those in the hostage dilemma, $\chi^2(1) = 13.76$, $p < .001$. About 82% of the participants (47 out of 57) in the car dilemma indicated that killing the individual was morally better than letting five people die, compared to 47% of the participants (26 out of 55) in the hostage dilemma. Figure 1 shows the means of participants' probability judgments in two dilemmas.

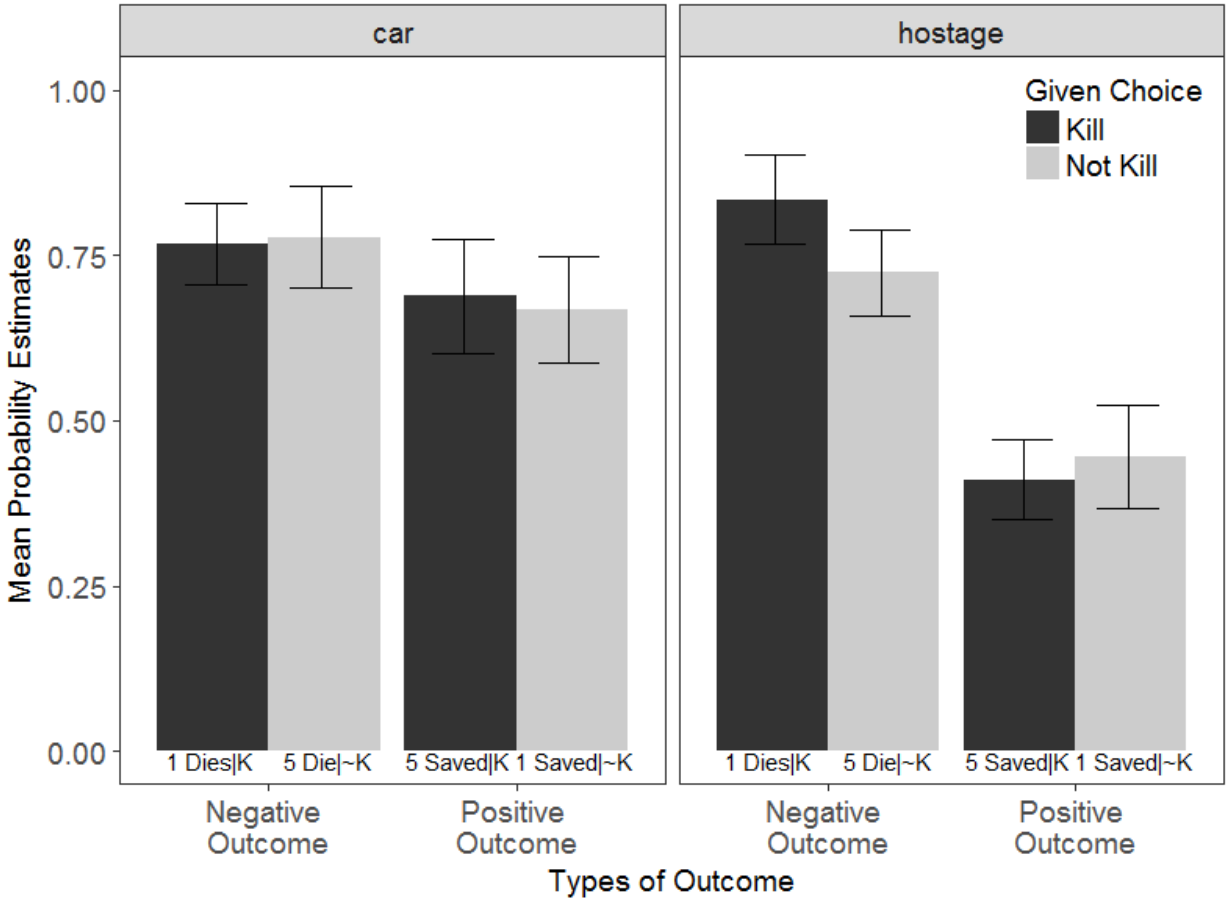


Figure 1. Mean probability judgments for the outcomes given the two alternatives between participants in the car dilemma and those in the hostage dilemma.

We first tested H1 that participants' probability judgments would be significantly different between different dilemmas using the intendent sample t-tests. Participants in the car dilemma provided significantly higher P(PO|K) and P(PO|~*K*) than the participants in the hostage dilemma did, $t(110) = 5.22$ and 4.00, respectively, *p*s <. 001. Participants' estimates for P(NO|K) and P(NO|~K) did not significantly differ between the two dilemmas, $t(110) = -1.49$ and 1.07, $p = .139$ and .287, respectively. We then compared the probability judgments between the two given choices using the paired t-tests. It was found that the mean perceived positive outcome probabilities were not significantly different between *K* and ~*K* in either the car or the hostage dilemmas. $|t| < 1$, *p*s> .370. The mean perceived negative outcome probabilities were also not significantly different between *K* and ~*K* in the car dilemma, $t(56) = -0.27$, $p = .786$. However, the perceived negative outcome probabilities given *K* were significantly higher than those given ~*K* in the hostage dilemma, $t(54) = 2.70$, $p = .009$.

Association between moral choices and probability judgments

Logistic regression was used to examine the associations between participants' moral choices and their probability judgments. The four probability judgments and dilemma types were used to predict participants' moral choices. Likelihood ratio tests suggest that there were no significant interaction effects between dilemma type and any of the four probability judgments, *p*s > .059.

Table 1 shows the estimation results of the final model. Subjects were more likely to choose *K* when their perceived probability of the positive outcome given *K* (*P*(PO|*K*)) was higher, $b = 1.28$, $p = .002$, or when their perceived probability of the negative outcome given ~*K* (*P*(NO|~*K*)) was higher, $b = 1.56$, $p < .001$. The perceived positive outcome given ~*K* (*P*(PO|~*K*)) reduced the likelihood of choosing *K*, while the perceived negative outcome given *K* (*P*(NO|*K*)) did not have significant influence on subjects' moral decisions.

Table 1: Logistic regression model for moral choices by the types of dilemma, the four probability judgments and order of task

	r	*b*	*SE*	*t*	*p*
Intercept		1.31	0.37	3.54	<.001
P(NO|*K*)	0.00	-0.17	0.28	-0.61	.541
P(PO|*K*)	0.36	1.28	0.42	3.04	.002
P(NO|~*K*)	0.48	1.56	0.37	4.23	<. 001
P(PO|~*K*)	0.00	-0.71	0.32	-2.23	.026
Dilemma (car)		1.01	0.33	3.07	.002

Note. *P*(PO|*K*): the probability that five people would survive given that *K* is chosen; *P*(NO|*K*): the probability that one person would die given that *K* is chosen; *P*(NO|~*K*): the probability that five people would die given that ~*K* is chosen; *P*(PO|~*K*): the

probability that one person would survive given that ~*K* is chosen. The four probabilities were standardized. *r* is the Pearson's correlations between the probability estimates and the dummy coded choices.

Discussion

Our study examined how participants perceived the outcome probabilities in different dilemmas. Overall, participants in the car dilemma provided probability judgments for the positive outcomes that were significantly higher than those in the hostage dilemma. With regards to the negative outcomes, participants in the hostage dilemma perceived a higher likelihood that the individual would die if *K* (killing one person) had been chosen, than that the five people would die if ~*K* (not killing) had been chosen. It is clear that participants perceived the utility given *K* in the car dilemma to be higher than for the hostage dilemma. This may explain why there was a higher proportion of participants in the car dilemma that chose *K* than in the hostage dilemma.

We also examined the association between participants' probability judgments and their moral decisions. The probabilities of outcomes that involved the five people had a greater association with participants' moral choices for both the car and hostage dilemmas than those that involved the individual. Participants were more likely to choose *K* if they perceived the five people had a higher chance of surviving ($P(\text{PO}|K)$) with the sacrifice of the individual, or that the five people had a higher chance of dying if they did nothing. These results suggested that participants preferred ~*K* more in the hostage dilemma than in the car dilemma, which might be due to a perception of higher loss and lower gain of *K* in the car dilemma. This also implied that participants were likely applying consequentialist reasoning even though they chose ~*K*.

General Discussion

In the present paper, we explored whether the discrepancy in participants' moral choices between two different dilemmas can be contributed to their perceptions of the outcome probabilities. It was found that participants' perceptions of outcome probabilities were significantly different between the two dilemmas utilized in the current study. Participants perceived the positive outcome was less likely to occur in the hostage dilemma than in the car dilemma. This also implies that participants perceived that the expected utility of each of the two choices (kill versus not kill) can differ across different dilemmas due to different perceptions of the outcome probabilities.

Results also showed that participants' perceptions of the outcome probabilities significantly predicted their moral choices. Participants were less likely to choose a choice if they perceived higher probability of the negative outcome given that choice. The pattern is consistent with a consequentialist approach to moral judgments, where participants prefer a choice that minimizes negative outcomes. Furthermore, this tendency (i.e., avoiding a choice when the negative outcome was more likely) was similar between the two dilemmas in the present study. This implies that the tendency to endorse a consequentialist approach among participants may not depend on the features of the dilemma such as being personal or impersonal.

Perhaps the most interesting finding was that the mean probabilities of the four outcomes were all well below 100%, even though the instruction stated that these outcomes "*will/will not*" occur given an action. Participants seemed to refuse to believe that one or five individual(s) will or will not die if ~*K* or *K* was taken. This hinged on a phenomenon called "failure to accept the task", which was first reported by Henle and Micheal (1956) and studied later by Richrer (1957). It was found that participants evaluated the content of the conclusion rather than the logical form of the argument when being asked to do a logical task. Richer (1957) suggested that the "failure to accept the task" might be due to "a general failure to grasp the concept of "logical validity" or one's "specific inability to differentiate 'logical validity' from other attributes of syllogisms" (p. 341). Consequently, participants were not performing "logical reasoning" as expected by the experimenter. In moral reasoning tasks, participants seem to make judgments based on their beliefs of the outcome possibilities instead of the information provided in the task instruction. Participants may experience difficulty in differentiating a hypothetical moral scenario from a real-world incident. They make their moral choices based on what they perceive as reasonable or consistent with their perceptions about the reality. This finding may highlight a main weakness of the moral reasoning studies that rely on descriptive tasks, where participants are likely to feel the scenarios are unreal.

Finally, the investigation of the expected values in the outcomes revealed that many participants chose ~*K* even when their expected outcome values of ~*K* was smaller than the ones of *K*. This suggests that factors other than outcome probabilities could affect moral choices in different dilemmas. One possible explanation is that participants' perceived "protected values" differ between the two dilemmas (Baron & Spranca, 1997). "Protected values" refer to the values that are against the trade-off in other type (economic/outcome) of values between the two choices. Taking a choice (usually an action such as killing) is at the cost of the "protected value" in addition to the death of the individual(s) given that choice. Participants might endorse higher protected values against killing for the hostage dilemma than they did for the car dilemma. Nevertheless, more research is needed to examine how moral decisions can be the joint product of both perceived outcome probabilities and protected values.

Overall, the present studies indicated that choosing ~*K* does not necessarily entail that people engage in deontological reasoning, and choosing *K* does not necessarily entail that people engage in consequentialist reasoning. Preferring a choice (e.g., *K*) in dilemma *A* (e.g., switch) more than dilemma *B* (e.g., footbridge) could be because participants perceive the positive (or negative) outcome given that choice is more (or less) likely in *A* than in *B*. Participants may not accept the instructions in the descriptive moral reasoning task and may not conduct moral reasoning with the information provided as per experimenters' expectations. Without controlling for the equivalence of the outcome probabilities perceived by participants, it could be inadequate to derive conclusions such as that the discrepancy in moral choices between two dilemmas is because one dilemma induces higher emotional arousal than the other. Robustness of current findings, however, should be examined in future research with the inclusion of more variants of moral dilemmas.