Commonsense psychology in human infants and machines: A Reproducibility Study

Peng, Xu[1], Yingjie, Zhou[1], Kaiyan, Shu[1], Yilin, Wang[1], & Yuqiang, Zhu[1]

[1]Nanjing Normal University

Introduction

Commonsense AI (AI) is an emerging direction of AI research that aims to enable AI systems to understand and reason about natural language and to commonsense knowledge as humans do.

In order to build commonsense AI, it is crucial to address the differences between humans and AI. One of the challenges is to decide the starting point. If the goal of common-sense AI is to build commonsense thinking ability like adults, then AI may need to start with the core competencies of infants, just as adults do.

Over the past few decades, basic research on infant commonsense psychology (i.e., infants' understanding of the intentions, goals, preferences, and rationality behind agent actions) has shown that infants attribute goals to the agent and expect the agent to pursue them in a reasonable and efficient manner. The prediction that infants possess commonsense psychology is fundamental to human social intelligence and can provide additional information for building commonsense AI. However, machine learning algorithms typically lack such predictions and instead predict actions directly, thus lacking flexibility for new environments and situations.

Thus, a comprehensive framework is needed to describe infants' knowledge of agent that allows comparable task outcomes between infants and machines. Such a framework could inform theories of infant knowledge and future human-like artificial intelligence.

The present study provides such a comprehensive framework for testing infants' commonsense psychology knowledge by assessing infants' performance on the Baby Intuitions Benchmark (BIB). The BIB contains six tasks that probe the psychology of commonsense and are applicable to both computational models and infants.

The underlying assumption of the BIB is that infants have an intrinsic, universally applicable cognitive ability to respond rapidly and unconsciously to external stimuli. Such responses are not subject to a process of verbal or logical reasoning, but are based directly on sensation and emotion. To test this hypothesis, the BIB typically uses a series of visual, auditory, or tactile stimuli and records the infant's reaction time and behavioral performance. Researchers can analyze this data to determine whether infants are able to recognize specific stimuli and how they respond.

Therefore, the present study is divided into two parts. Part one contains two experiments, which collected infants' responses to six tasks of the BIB, aim to provide preliminary evidence of infants' commonsense psychology and three attribution methods. Part two compares the performance of a state-of-the-art learning-driven neural network model on the BIB with that of infants to test whether infants' intelligence on the agent can be reflected in artificial intelligence. We focus on the infant experiment for replication.

Therefore, the present study was divided into two parts. The first part contains two experiments, which collected preliminary evidence supporting the infant's commonsense psychology and inferring three ways of attribution by observing the infant's performance in the BIB tasks. The second part tests whether infants' knowledge of agent can be reflected in AI by comparing a state-of-the-art learning-driven neural network model to infants' performance on the BIB tasks. In the repetition study, we mainly focus on the experiments in the first part.

Material and Method

The BIB uses the "Violation of Expectation" (VOE) paradigm, which is commonly used in infant experiment, and has been used in recent machine learning benchmarks focused on commonsense.

BIB's six tasks consisted of a series of short silent animated videos with simple visual effects: some basic shapes without human features (e.g., eyes and limbs). The experiment was divided into two phases: familiarization phase and test phase. In

familiarization phase, observer watched a series of videos with the aim of establishing an expectation. Videos in test phase had two outcomes. One was the expected outcome, which was different in perception from the familiarization phase but consistent in concept. The second was the unexpected outcome, which was similar in perception to the familiarization phase but conceptually inconsistent.

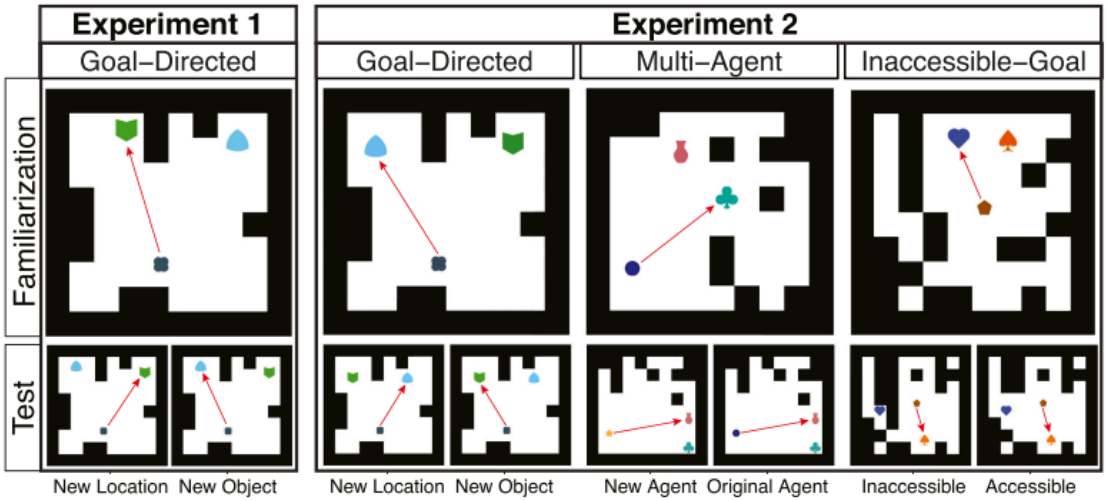
For each task, observers first saw eight familiarization trial videos in which an agent acts consistently in terms of its goals, rationality, or instrumentality. The exact make-up of the grid world and the movement of the agent may vary across trials, as described in the main text. One example still image per task from a familiarization trial video is shown here. Observers then saw expected and unexpected test trial videos (with the order of these trials varying for infants). Example still images of both test trial videos per task are shown here.

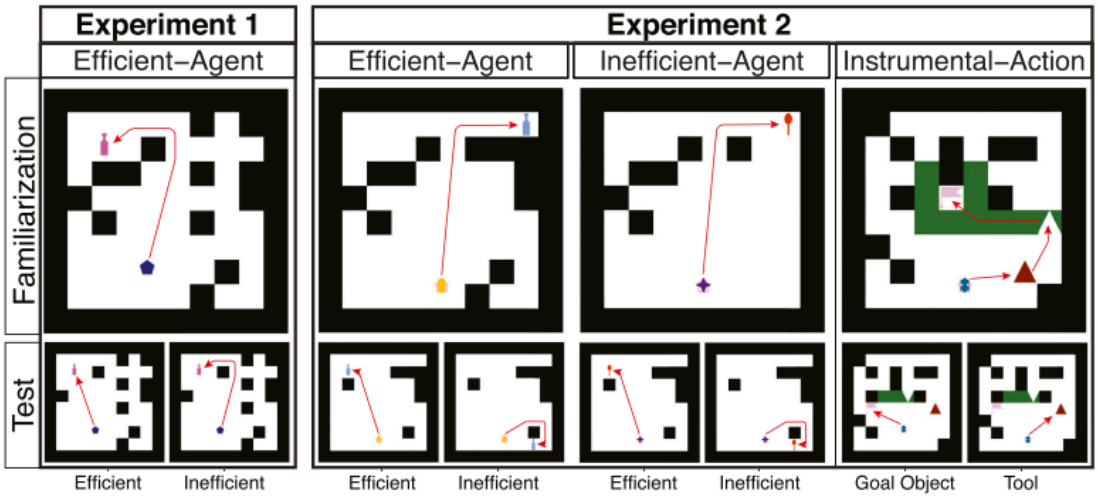
The first three tasks focus on an observer 's attribution of goals to agents' actions. The *Goal-Directed Task* captures the idea that agents' goals are directed towards objects, not locations. Observers watch an agent repeatedly move to the same one of two objects in approximately the same location in an unchanging grid world during familiarization. At test, observers may be more surprised when the agent moves to a new object in that grid world after the locations of the two objects switch. The *Multi-Agent Task* asks whether goals are specific to agents. Observers watch an agent move to the same one of two objects during familiarization in a changing grid world, with both objects appearing in varying locations. At test, observers may be more surprised when the original agent versus a new agent moves to a new object. The *Inaccessible-Goal Task* asks whether agents might form new goals when their existing goals become unattainable. Observers watch an agent move to the same one of two objects during familiarization in a changing grid world, with both objects appearing in varying locations. At test, the grid world changes again such that the agent's goal object becomes physically inaccessible. Observers may be more surprised when the agent moves to a new object when its prior goal object is accessible versus inaccessible.

The next two tasks focus on an observer's attribution of rationality to agents' actions. The *Efficient-Agent Task* captures the idea that agents act rationally to achieve

goals. Observers watch an agent move to an object efficiently around obstacles in an unchanging grid world during familiarization. At test, the object appears in a location that it had appeared during familiarization, but the grid world has changed such that the obstacles that blocked the object are gone or have been replaced with different obstacles. Observers may be more surprised when the agent moves along a familiar but now inefficient path to the object. The *Inefficient-Agent Task* asks what expectations observers have about agents who initially move inefficiently in a changing grid world. During familiarization, observers watch an agent move along the same paths to an object as the agent in the *Efficient-Agent Task*, but this time there are no obstacles in the agent's way, so the agent 's movements to the object are inefficient. At test, the environment changes as in the *Efficient Agent Task*. Observers may either be more surprised when the agent continues to move inefficiently to the object or may have no expectations about whether that agent will move efficiently or inefficiently to the object.

The last task focuses on an observer's attribution of instrumentality to agents' actions. The *Instrumental-Action Task* captures the idea that agents should only take instrumental actions when necessary. During familiarization, observers watch an agent move first to a key, which it uses to remove a barrier around an object in varying locations, and then to that object. At test, observers may be more surprised when the agent continues to move to the key, instead of directly to the object, when the barrier is no longer blocking the object.





Participants

In Experiment 1, typically developing 11-month-old infants ($N = 26$, $M_{age} = 11.13$ months, Range = 10.42 months - 11.83 months; 12 girls) born at ≥37 weeks gestational age were included. They completed the *Goal-Directed Task*, the *Efficient-Agent Task*, or both, with half of the infants receiving each task first, totaling $N = 48$ individual testing sessions and $N = 24$ sessions per task. An additional four sessions were excluded because infants did not complete the session.

In Experiment 2, typically developing 11-month-old infants ($N = 58$, $M_{age} = 11.06$ months, Range = 10.50 months - 11.50 months; 31 girls) born at ≥37 weeks gestational age were included. Each infant completed at least one of BIB's tasks, totaling $N = 288$ individual testing sessions.

Following our preregistration, data collection stopped when 32 infants (*M*age = 11.09 months, Range = 10.50 months – 11.50 months; 17 girls) completed all six of BIB's tasks. Tasks were presented in a semi randomized order using 32 fixed orders that averaged to each task being presented 5.33 times in each ordinal position (range: 4-7 times).

The final sample sizes for each task were: *Goal-Directed Task*, $N = 48$; *Multi-Agent Task*, $N = 49$; *Inaccessible-Goal Task*, $N = 47$; *Efficient Agent Task*, $N = 47$; *Inefficient-Agent Task*, $N = 49$; *Instrumental-Action Task*, $N = 48$.

An additional 37 sessions were excluded because of preregistered exclusion

criteria, including: looking time < 1.5s to least one test trial and/or two familiarization trials with or without the infant completing the session (16); poor video quality and/or technical failure (18); and caretaker interference (3). An additional two sessions were excluded post hoc for extreme values (> 40s) to one test outcome, which could artificially inflate the calculation of the sample's variance. These extreme values were identified through examination of a histogram of the raw looking times across all of the sessions and across all of the tasks by two researchers masked to the task and outcome represented by each value. Exclusions were consistent across tasks: *Goal-Directed Task*, 5; *Multi-Agent Task*, 6; *Inaccessible-Goal Task*, 9; *Efficient-Agent Task*, 7; *Inefficient-Agent Task*, 5; *Instrumental-Goal Task,* 7. The total exclusion rate was 11.9%.

Procedure

Infants were tested online on Zoom. In the first ten minutes of the first testing session, the experimenter explained to caretakers the instructions for setting up their device and for positioning the infant in front of the screen. Caretakers are required to close their eyes and not communicate with the infant during the stimuli presentation. The experimenter, masked to what trial was being presented and the order of the test trials, coded infants' looking to the stimuli live from the start of each video and controlled the progression of stimuli using PyHab (Kominsky, 2019) and *slides.com*. Each trial video was preceded by a 5s attention grabber (a swirling blob accompanied by a chiming sound, centered on the screen) to focus the infant's attention to the screen, and each video froze after the agent reached an object. The last frame of the video remained on the screen until infants looked away for 2s consecutively or for a maximum of 60s. Testing sessions were recorded through the Zoom recording function, capturing both the infant's face and the screen presenting the stimuli.

Result

Infants' performance on Experiment 1's two tasks is displayed in Fig. 2. Infants' looking time varied by task, with longer looking to the Efficient-Agent versus Goal-

Directed Task (F (1, 71) = 9.34, p = .003), reflecting the longer test-trial lengths in the Efficient-Agent Task (see SI). Overall, infants looked longer to the unexpected versus expected outcomes (F (1, 66) = 11.34, p = .001), and there was no task by outcome interaction (F (1, 66) = 0.30, p = .585). Infants were surprised (looked longer) when an agent moved to a new object in the Goal-Directed Task (F (1, 23) = 4.73, p = .040), and they were surprised when an efficient agent later took an inefficient path to an object in the Efficient-Agent Task (F (1, 23) = 2.60, p = .016).

Infants' performance on Experiment 2's six tasks is also displayed in Fig. 2. Infants' looking time varied by task (F (5, 341) = 2.78, p = .018), reflecting the different test-trial lengths of the different tasks. Overall, infants did not look longer to unexpected versus expected outcomes (F (1, 341) = 2.27, p = .133), but a task by outcome interaction suggested that different tasks elicited different patterns of infants' looking (F (5, 341) = 2.23, p = .051).

We first considered infants' performance on Experiment 2's three tasks that focused on goal attribution: the *Goal-Directed*; *Multi-Agent*; and *Inaccessible-Goal Tasks*. First, consistent with the results in Experiment 1, infants were surprised when an agent moved to a new object in the *Goal Directed Task* (F (1, 47) = 4.09, p = .049). Infants presented with a new agent in the *Multi-Agent Task*, however, did not show a difference in surprise when that agent versus the original agent moved to a new object (F (1, 48) = 3.41, p = .071; with longer looking times to the expected outcome). Infants in the *Inaccessible-Goal Task* also did not show a difference in surprise when an agent moved to a new object when its goal object was accessible versus inaccessible (F (1, 46) = 0.02, p = .891).

We next considered infants' performance on the two tasks that focused on rationality attribution: the *Efficient-Agent* and *Inefficient Agent Tasks*. First, consistent with the results in Experiment 1, infants were surprised when an efficient agent later took an inefficient path to an object in the *Efficient-Agent Task* (F (1, 46) = 7.72, p = .008). Infants in the *Inefficient-Agent Task* did not show a difference in surprise when an inefficient agent continued to move inefficiently to an object at test (F (1, 48) = 2.51, p = .119). But, when comparing infants' performance in the *Efficient-Agent* and

Inefficient-Agent Tasks directly, there was no significant task by outcome interaction (F (1, 132) = 0.49, p = .484): We did not find evidence that infants' surprise at the inefficient agent's later inefficient action was different from their surprise at the efficient agent's later inefficient action.

Finally, we considered infants instrumentality attribution through their performance on the *Instrumental-Action Task*. Infants did not show a difference in surprise when the agent moved to the tool as opposed to its goal object when the tool was no longer needed to achieve the goal (F (1, 47) = 0.03, p = .853).

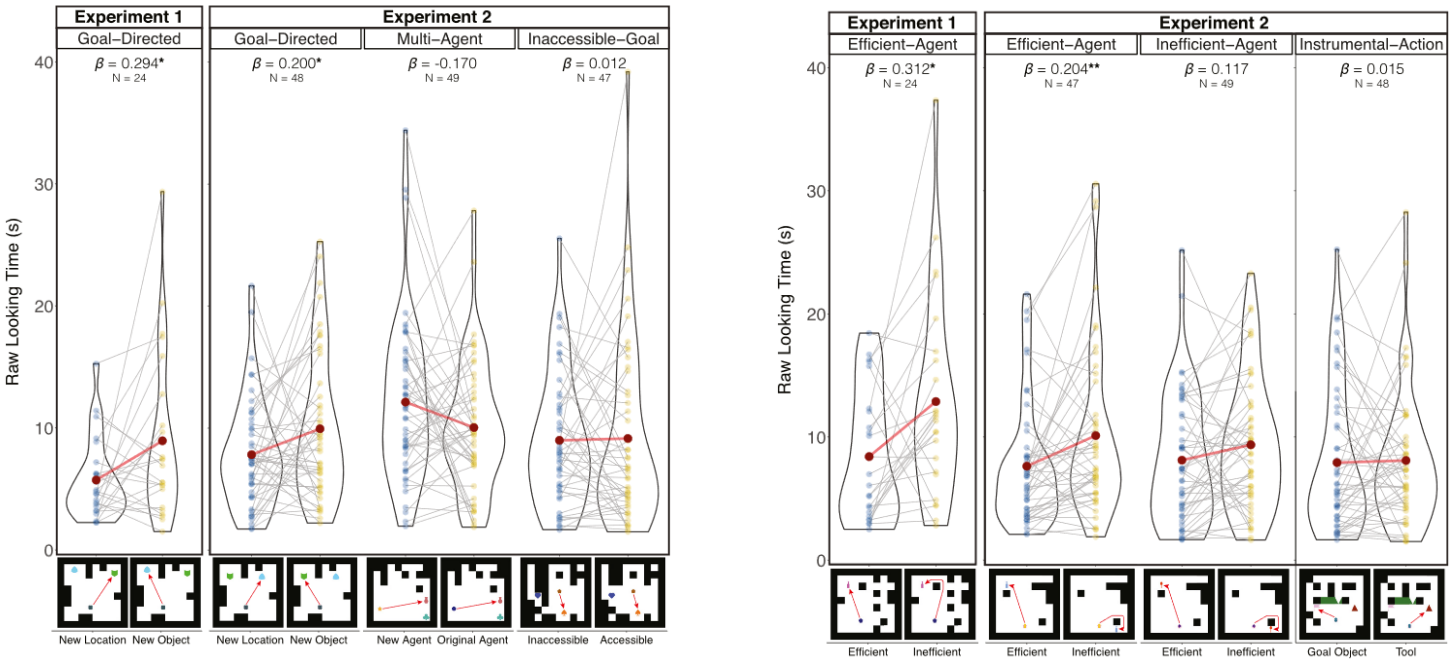


Fig. 2. Infants' raw looking times to the two outcomes in each of BIB's tasks in Experiments 1 & 2. Gray lines connect the individual looking times (represented by blue and yellow dots) of each infant to each outcome. Red dots connected by red lines indicate the mean looking times to each outcome for each task. Beta coefficients are effects sizes in terms of standard deviations, and statistical analyses are reported in the main text (*p < .05, **p < .01). (For interpretation of the references to color in this figure legend, please see the online version.)

Reproducibility Study

Repeat idea

The reproduction process is as follows: first, define the relevant functions. Then read

the data that needs to be used. After the reading is completed, the data cleaning work is carried out, with the main purpose of clearing extreme data with fixation time less than 1 second and more than 40 seconds (situations such as incomplete experimental tasks, technical failures, main trial issues, poor video quality, and caregiver interference have been eliminated). Then establish an HLM model and perform a Type 3 Wald test on the data. Finally, output the results.

Procedure

Set work path

```
WD <-  here::here()
getwd()
```

Load packages

```
if (!requireNamespace("pacman", quietly = TRUE)) {
  install.packages("pacman") }
pacman::p_load("here", "bruceR", "reticulate")

use_python('C:\\Users\\徐鹏\\AppData\\Local\\Programs\\Python\\Python310\\python.exe')
```

Define functions

```
P_summary<- function(result){
  cat("\033[1;30m
Type III Analysis of Variance Table with Satterthwaite's method: \033[0m\n")
  print_table(result)
}

P_cleaning<- function(data){
  if('Task' %in% colnames(data)){
    setDT(data)
    data<- data[Looking >= 1 & Looking < 40]
    data = select(data,c('Participant','Task','Outcome','Looking'))
  }else{
    setDT(data)
    data<- data[Looking >= 1 & Looking < 40]
    data = select(data,c('Participant','Outcome','Looking'))
  }
}

P_wider<- function(data,id,dv,idv){
  v = c(id,dv,idv)
  data = select(data,v)
  pivot_wider(data = data,
              names_from = idv,
              values_from = dv)
}

P_describe<- function(data){
  data_summary = P_wider(data, id = 'Participant', dv = 'Looking', idv = 'Outcome')
  data_summary = select(data_summary, c('Unexpected','Expected'))
  Describe(data_summary,plot = TRUE,upper.triangle = TRUE)
}
```

P_ The summary function is used to present the results of Type 3 Wald tests in one click. P_ Cleaning is used for one-step cleaning and filtering of data. P_ Wider is used to convert the original data into wider data in one step. P_ Describe is used for one-step descriptive statistics.

Experiment 1

Loading and cleaning data

There are three data in Experiment 1, which are effective_ Agent, goal_ Directed data and an omnibus comprehensive data.

```
efficient_agent<- read.csv('data/Exp1/efficient_agent.csv', encoding = 'UTF-8', header = TRUE)%>%
  P_cleaning()
goal_directed<- read.csv('data/Exp1/goal_directed.csv', encoding =  'UTF-8', header = TRUE)%>%
  P_cleaning()
omnibus<- read.csv('data/Exp1/omnibus.csv', encoding = 'UTF-8', header = TRUE)%>%
  P_cleaning()
```

Comprehensive analysis of Omnibus

```
model1<- lmer(Looking ~ Outcome * Task + (1|Participant), data = omnibus) #建立HLM模型
result1<- anova(model1) #对建立的模型进行 Type 3 tests
P_summary(result1) #呈现最终的结果
```

```
Type III Analysis of Variance Table with Satterthwaite's method:
             Sum Sq Mean Sq NumDF  DenDF      F     p
Outcome      360.248 360.248 1.000 65.530 11.336  .001 **
Task         296.902 296.902 1.000 71.195  9.343  .003 **
Outcome:Task   9.566   9.566 1.000 65.530  0.301  .585
```

The analysis results are consistent with the original text: the ratio of efficient agent to *Goal oriented task* (F (1, 71) = 9.34, p=. 003), the ratio to expected result (F (1, 66) = 11.34, p=. 001), no task to result interaction (F (2, 66) = 0.30, p=. 585), G*oal oriented task* (F (3, 23) = 4.73, p=. 040), *efficient agent task* (F = 2.60, p=. 016).

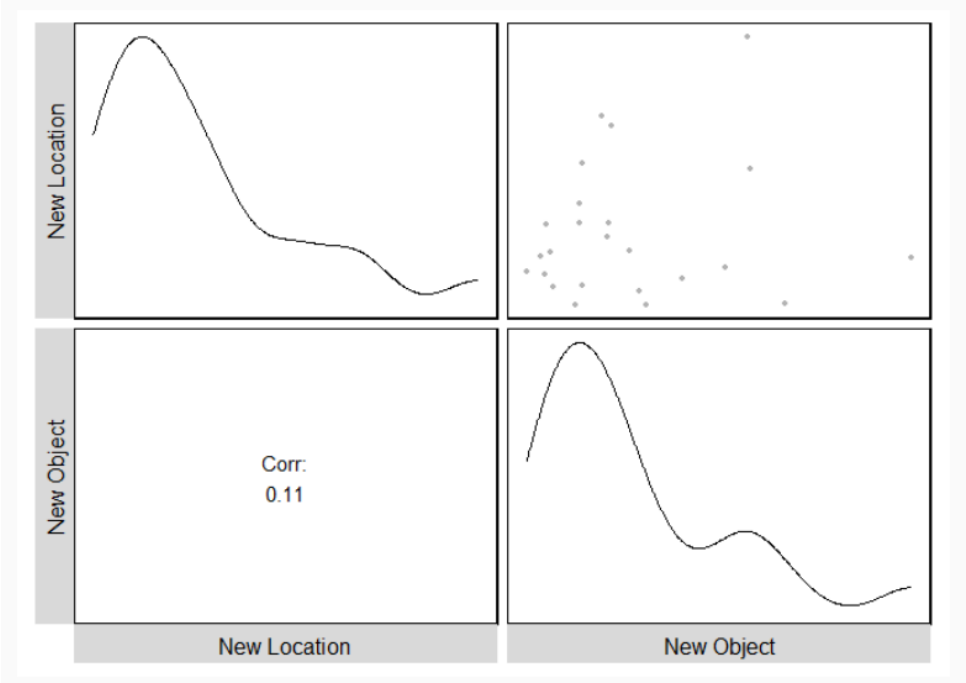
Omnibus results visualization

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.ticker as ticker
```

```
data = pd.read_csv('E:\R course\omnibus.csv') #读取数据
plt.rcParams['font.size'] = '12' #设置字号
sns.set(style = 'ticks') #设置主题
g = sns.FacetGrid(data, col='Task') #设置绘图网格
g.map_dataframe(sns.violinplot, #小提琴函数
                x = 'Outcome', #自变量
                y = 'Looking', #因变量
                scale='count', #根据样本量调节图形宽度
                saturation=1, #饱和度
                inner='box', #中间呈现微型箱型图
                linewidth=1, #边缘曲线宽度
                split=True, #分开呈现
                bw = 0.5, #核密度估计值，越大越光滑
                palette='Pastel2' #设置颜色
                )
```

The *Goal-Directed Task*

```
goal_directed_summary<- P_wider(data = goal_directed, id = 'Participant',
dv ='Looking',idv = 'Outcome')%>%
  select(.,c('New Location','New Object'))%>%
  Describe(., plot = TRUE,upper.triangle = TRUE) #描述性统计
```




```
model2<- lmer(Looking ~ Outcome + (1|Participant), data = goal_directed) #建立HLM模型
result2<- anova(model2) #对建立的模型进行 Type 3 tests
P_summary(result2) #呈现最终的结果
```

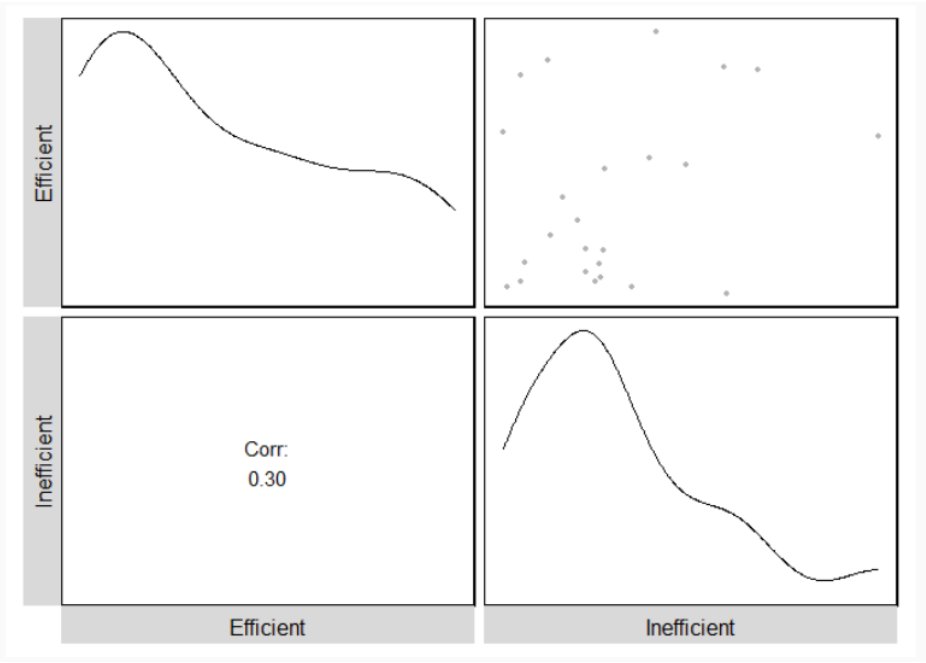
```
Descriptive Statistics:

                 N Mean   SD | Median  Min   Max Skewness Kurtosis
New Location  24 5.26 3.32 |   4.23 1.75 14.87     1.23     0.87
New Object    24 8.50 6.87 |   6.70 1.00 29.08     1.30     1.13

Type III Analysis of Variance Table with Satterthwaite's method:
          Sum Sq Mean Sq NumDF  DenDF     F     p
Outcome  126.204 126.204 1.000 23.000 4.731  .040 *
```

The analysis results are consistent with the original text: the *Goal Directed Task* (*F* (1, 23) = 4.73, *p*=. 040).

The *Efficient-Agent Task*

```
efficient_agent_summary<- P_wider(data = efficient_agent, id =
'Participant', dv ='Looking',idv = 'Outcome')%>%
  select(.,c('Efficient','Inefficient'))%>%
```

```
Describe(., plot = TRUE,upper.triangle = TRUE) #描述性统计
```




```
model3<- lmer(Looking ~ Outcome + (1|Participant), data = efficient_agent) #建立HLM模型
result3<- anova(model3) #对建立的模型进行 Type 3 tests
P_summary(result3) #呈现最终的结果
```

```
Descriptive Statistics:
──────────────────────────────────────────────────────────────────
              N  Mean   SD | Median  Min   Max Skewness Kurtosis
──────────────────────────────────────────────────────────────────
Efficient    24  7.96 5.31 |   6.06 2.01 18.07     0.52    -1.28
Inefficient  24 12.47 8.37 |  11.04 2.32 37.13     1.12     0.95
──────────────────────────────────────────────────────────────────

Type III Analysis of Variance Table with Satterthwaite's method:
──────────────────────────────────────────────────────────────
          Sum Sq Mean Sq NumDF  DenDF     F     p
──────────────────────────────────────────────────────────────
Outcome  243.610 243.610 1.000 23.000 6.773  .016 *
──────────────────────────────────────────────────────────────
```

The analysis results differ from the original text in terms of F-value: the *Efficient Agent Task* (F (1, 23) = 2.60, p=. 016). By checking the author's publicly available R code, it can be seen that the author has filled in the wrong F value here.

Experiment 2

Ominibus

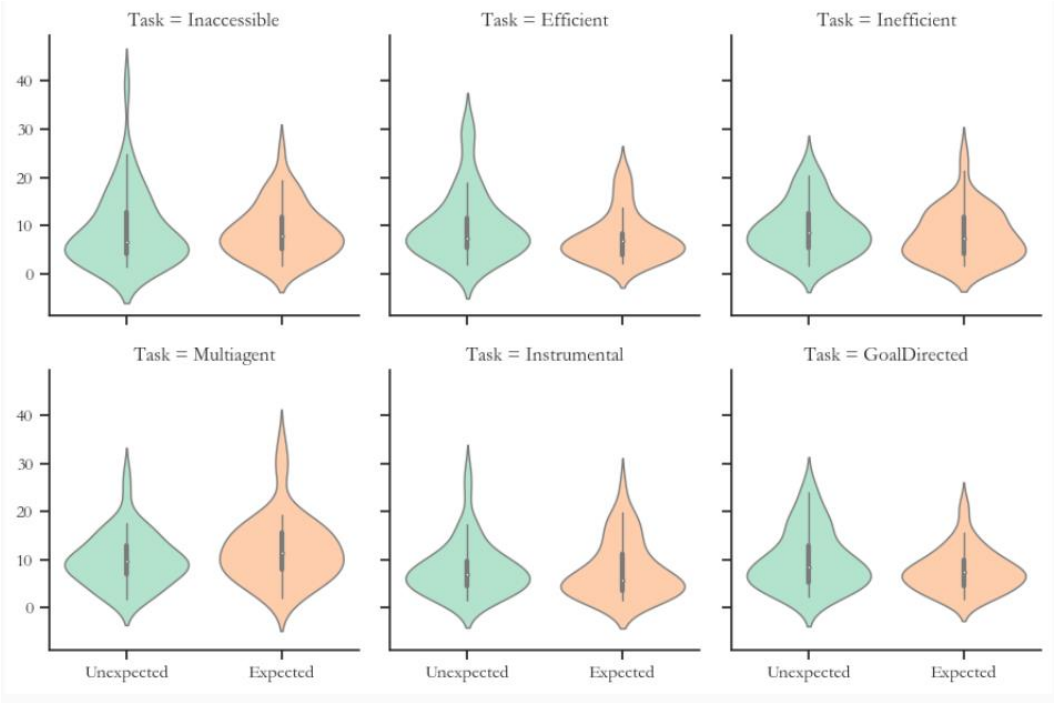
```
Exp2_omnibus<- read.csv('data/Exp2/Data from all participating infants/omnibus.csv',
                        encoding = 'UTF-8',
                        header = TRUE) %>%
  P_cleaning() #读取并清理数据
model4<- lmer(Looking ~ Outcome * Task + (1|Participant), data = Exp2_omnibus) #建立HLM模型
result4<- anova(model4) #对建立的模型进行 Type 3 Wald tests
P_summary(result4) #呈现最终的结果
Exp2_omnibusTest<- read.csv('data/Exp2/Data from infants who completed all six tasks/omnibus.csv')
length(unique(Exp2_omnibus$Participant)) #检查复现过程中所用数据的被试数量，为58
length(unique(Exp2_omnibusTest$Participant))#检查研究者所用数据的被试数量，为32
```

```
Type III Analysis of Variance Table with Satterthwaite's method:
             Sum Sq Mean Sq NumDF   DenDF     F      p
Outcome      67.063  67.063 1.000 508.373 2.260   .133
Task        522.013 104.403 5.000 523.226 3.519   .004 **
Outcome:Task 335.124  67.025 5.000 508.373 2.259   .047 *
```

The results of the analysis are inconsistent with the original text: Infants' looking time varied by task (F (5, 341) = 2.78, p=. 018), unexpected versatile expected outputs (F (1, 341) = 2.27, p=. 133), a task by output interaction suggested that different tasks were elicited different patterns of Infants' looking (F (5, 341) = 2.23, p=. 051). By examining the publicly available code of the researchers, it can be concluded that the author mistakenly analyzed the pre-experimental data as formal data.

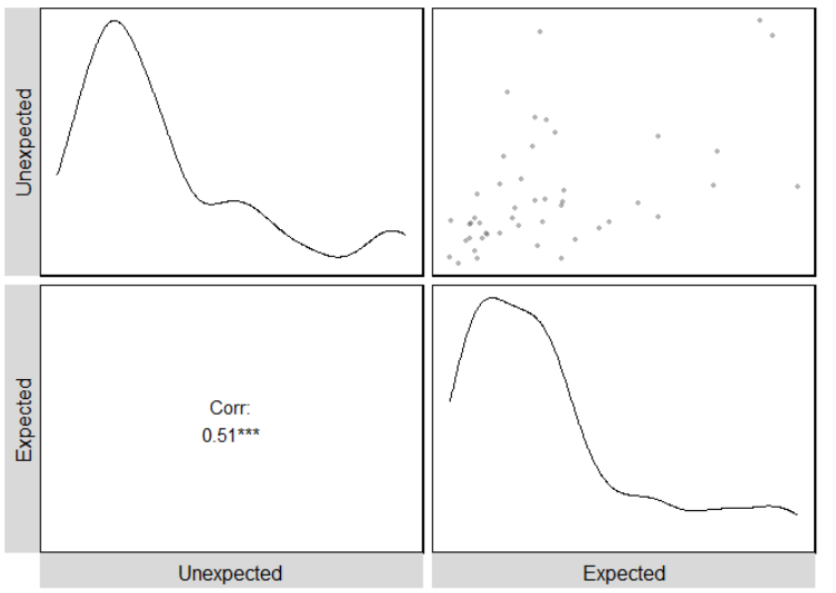
Omnibus results visualization

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import matplotlib.ticker as ticker
data = pd.read_csv('E:\R course\Data from all participating infants\omnibus.csv')

sns.set(font='STSong',font_scale=1,style = 'ticks')
g = sns.FacetGrid(data, col='Task',col_wrap=3)
g.map_dataframe(sns.violinplot,x = "Outcome",
                y = "Looking",
                scale = 'count',
                saturation = 1,
                inner = 'box',
                linewidth = 1,
                split = True,
                bw = 0.5,
                palette = 'Pastel2'
                )
```



The *Efficient Agent Task*

```
Exp2_efficient_agent<- read.csv('data/Exp2/Data from all participating infants/efficient_agent.csv',
                           encoding = 'UTF-8',
                           header = TRUE) %>%
  P_cleaning() #读取并清理数据
P_describe(Exp2_efficient_agent)
```




```
model5<- lmer(Looking ~ Outcome + (1|Participant), data = Exp2_efficient_agent) #建立HLM模型
result5<- anova(model5) #对建立的模型进行 Type 3 tests
P_summary(result5) #呈现最终的结果
```

Descriptive Statistics:

	N	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Unexpected	47	10.12	6.97	7.48	1.89	30.56	1.40	1.40
Expected	47	7.63	4.93	6.76	2.09	21.60	1.30	0.89

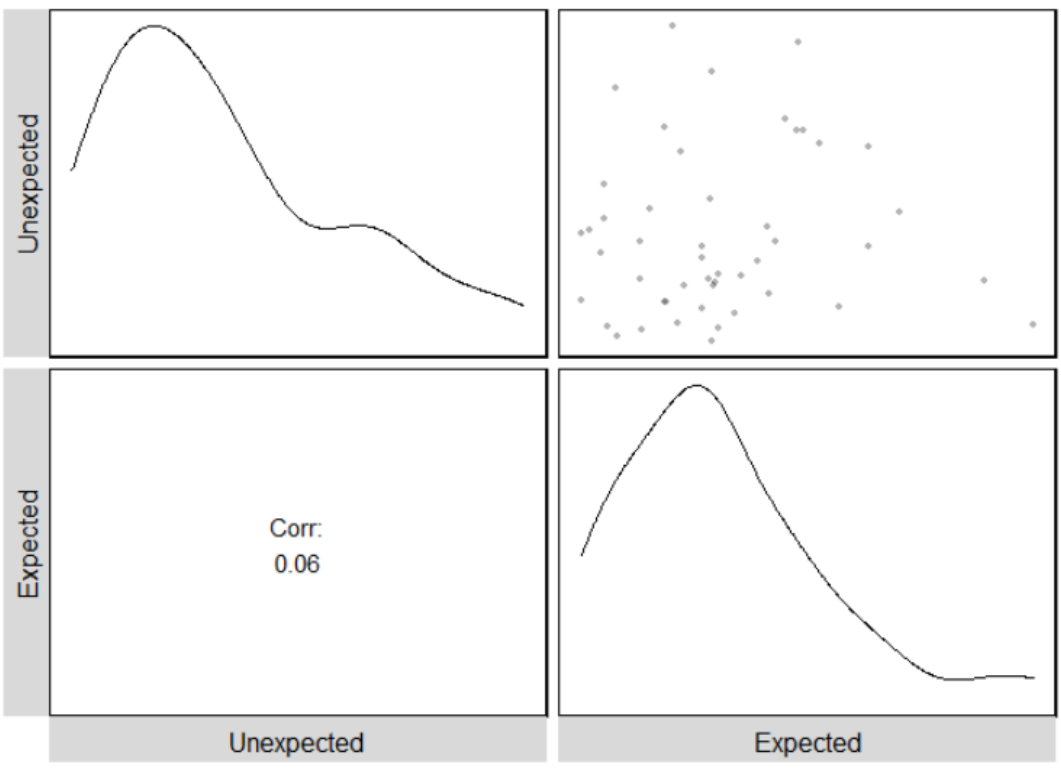
Type III Analysis of Variance Table with Satterthwaite's method:

	Sum Sq	Mean Sq	NumDF	DenDF	F	p
Outcome	145.538	145.538	1.000	46.000	7.719	.008 **

The analysis results are consistent with the original text: the *Efficient Agent Task* (*F* (1, 46) = 7.72, *p*=. 008).

The *Goal Directed Task*

```
Exp2_goal_directed<- read.csv('data/Exp2/Data from all participating infants/goal_directed.csv',
                              encoding = 'UTF-8',
                              header = TRUE) %>%
  P_cleaning() #读取并清理数据
P_describe(Exp2_goal_directed)
```



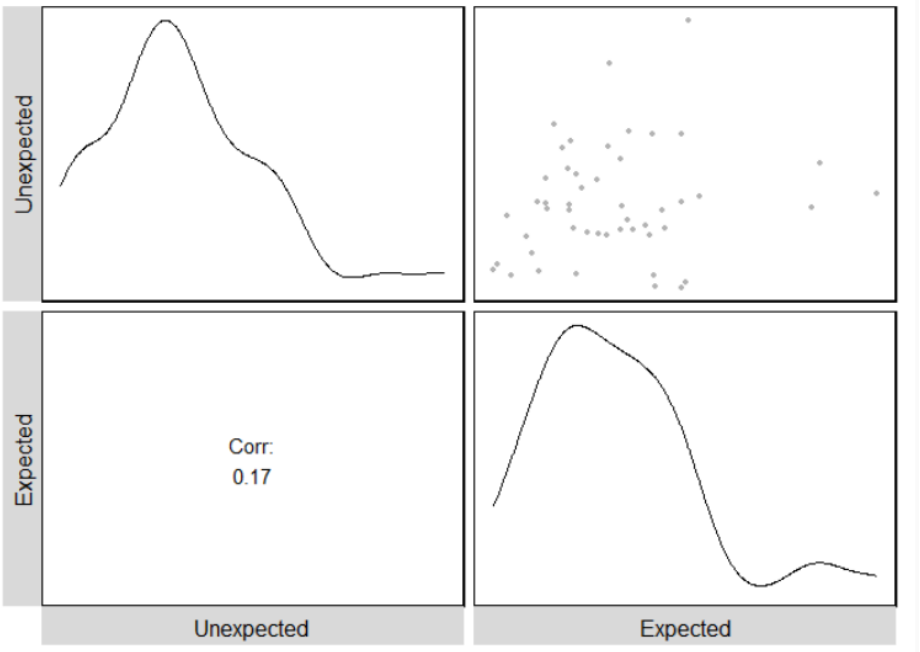
```
model6<- lmer(Looking ~ Outcome + (1|Participant), data = Exp2_goal_directed) #建立HLM模型
result6<- anova(model6) #对建立的模型进行 Type 3 tests
P_summary(result6) #呈现最终的结果
```

```
Descriptive Statistics:
──────────────────────────────────────────────────────────────
             N Mean   SD | Median  Min   Max Skewness Kurtosis
──────────────────────────────────────────────────────────────
Unexpected  48 9.95 6.02 |   8.53 2.22 25.28     0.84    -0.27
Expected    48 7.83 4.42 |   7.40 1.73 21.67     1.00     0.97
──────────────────────────────────────────────────────────────

Type III Analysis of Variance Table with Satterthwaite's method:
──────────────────────────────────────────────────────────────
          Sum Sq Mean Sq NumDF  DenDF     F     p
──────────────────────────────────────────────────────────────
Outcome  107.740 107.740 1.000 47.000 4.088  .049 *
──────────────────────────────────────────────────────────────
```

The analysis results are consistent with the original text: the *Goal Directed Task* (*F* (1, 47) = 4.09, *p*=. 049).

The *Multi-Agent Task*

```
Exp2_multi_agent<- read.csv('data/Exp2/Data from all participating infants/multi_agent.csv',
                            encoding = 'UTF-8',
                            header = TRUE) %>%
  P_cleaning() #读取并清理数据
P_describe(Exp2_multi_agent)
```



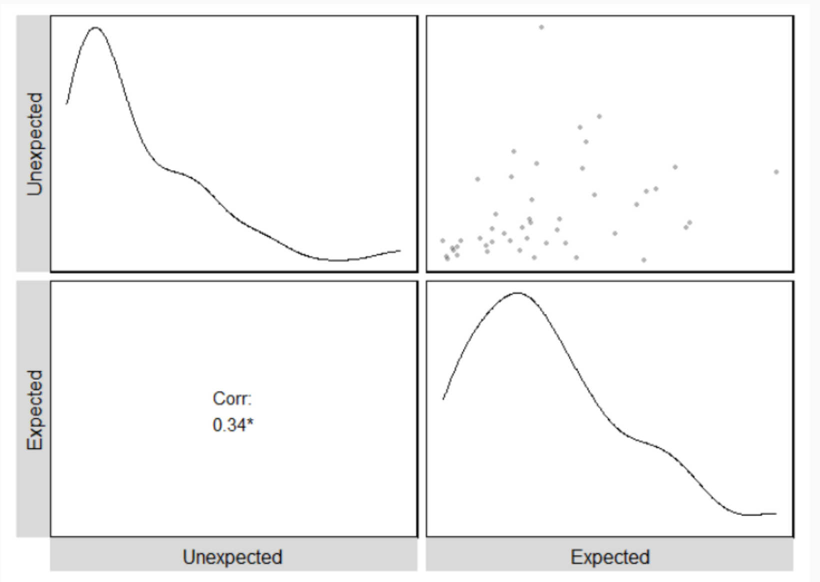
```
model7<- lmer(Looking ~ Outcome + (1|Participant), data = Exp2_multi_agent) #建立HLM模型
result7<- anova(model7) #对建立的模型进行 Type 3 tests
P_summary(result7) #呈现最终的结果
```

```
Descriptive Statistics:
─────────────────────────────────────────────────────────────
              N  Mean   SD | Median  Min   Max Skewness Kurtosis
─────────────────────────────────────────────────────────────
Unexpected   49 10.05 5.42 |   9.60 1.88 27.83     0.87     1.06
Expected     49 12.14 6.74 |  11.55 1.97 34.41     1.14     1.66
─────────────────────────────────────────────────────────────

Type III Analysis of Variance Table with Satterthwaite's method:
────────────────────────────────────────────────────
          Sum Sq Mean Sq NumDF  DenDF     F     p
────────────────────────────────────────────────────
Outcome  106.931 106.931 1.000 48.000 3.412  .071 .
────────────────────────────────────────────────────
```

The analysis results are consistent with the original text: the *Multi Agent Task* (F (1, 48) = 3.41, p=. 071).

The *Inaccessible-Agent Task*

```
Exp2_inaccessible_agent<- read.csv('data/Exp2/Data from all participating infants/inaccessible_goal.csv',
                                   encoding = 'UTF-8',
                                   header = TRUE) %>%
  P_cleaning() #读取并清理数据
P_describe(Exp2_inaccessible_agent)
```



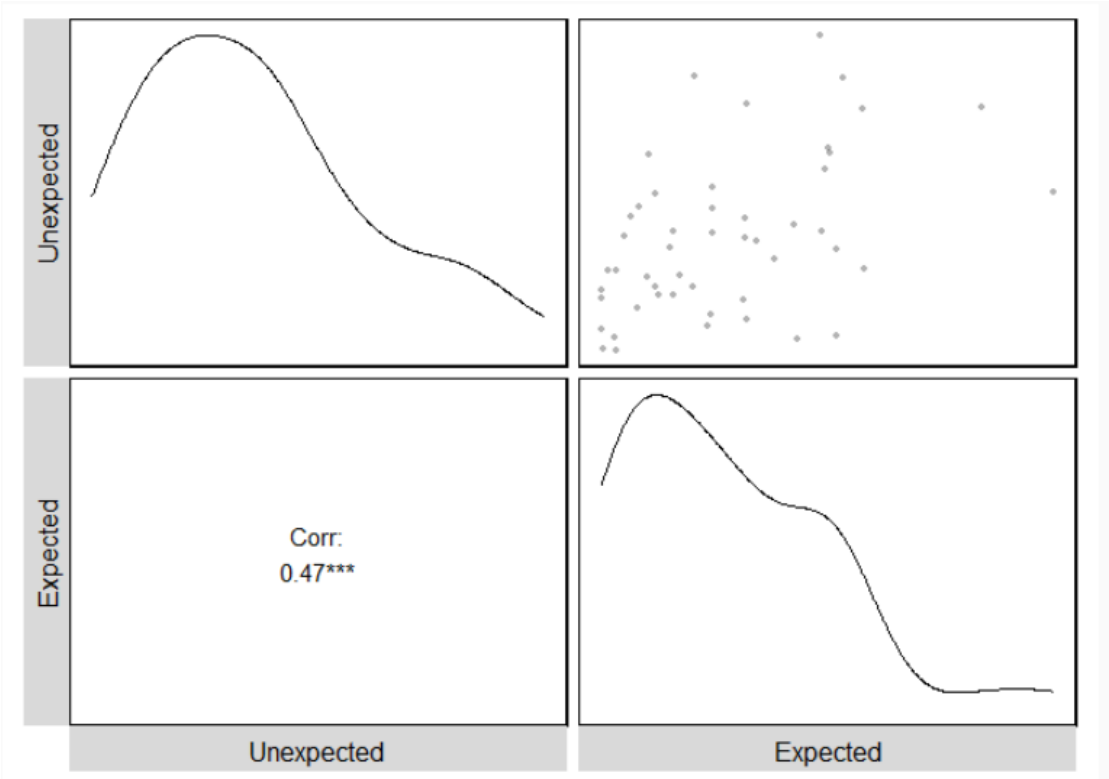
```r
model8<- lmer(Looking ~ Outcome + (1|Participant), data = Exp2_inaccessible_agent) #建立HLM模型
result8<- anova(model8) #对建立的模型进行 Type 3 tests
P_summary(result8) #呈现最终的结果
```

```
Descriptive Statistics:

                N Mean   SD | Median  Min   Max Skewness Kurtosis
Unexpected 47 9.15 7.55 |   6.66 1.50 39.21     1.67     3.41
Expected   47 9.00 5.46 |   7.96 1.68 25.54     0.83     0.23

Type III Analysis of Variance Table with Satterthwaite's method:

         Sum Sq Mean Sq NumDF  DenDF     F     p
Outcome   0.563   0.563 1.000 46.000 0.019  .890
```

The analysis results are consistent with the original text: the *Inaccessible Goal Task* (F (1, 46) = 0.02, p=. 891).

The *Inefficient-Agent Task*

```r
Exp2_inefficiet_agent<- read.csv('data/Exp2/Data from all participating infants/inefficient_agent.csv',
                                 encoding = 'UTF-8',
                                 header = TRUE) %>%
  P_cleaning() #读取并清理数据
P_describe(Exp2_inefficiet_agent)
```



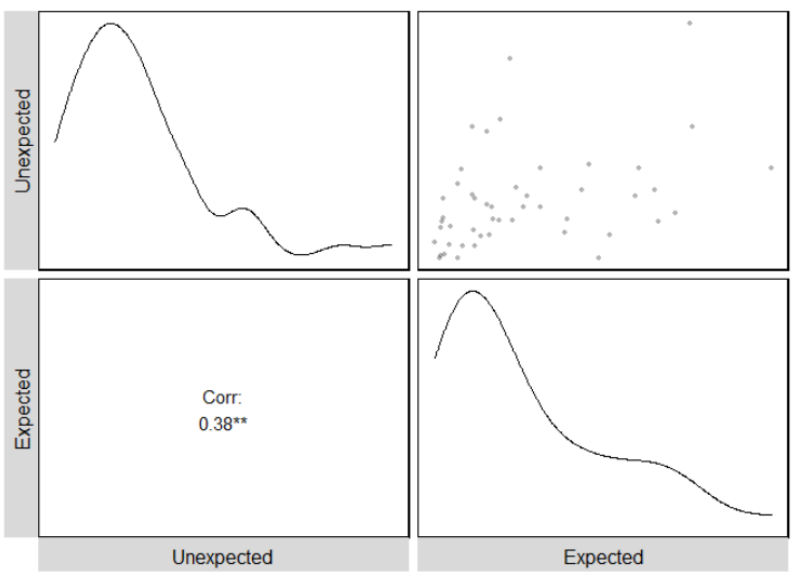
```
model9<- lmer(Looking ~ Outcome + (1|Participant), data = Exp2_inefficiet_agent) #建立HLM模型
result9<- anova(model9) #对建立的模型进行 Type 3 tests
P_summary(result9) #呈现最终的结果
```

```
Descriptive Statistics:
──────────────────────────────────────────────────────────────
             N Mean   SD | Median  Min   Max Skewness Kurtosis
──────────────────────────────────────────────────────────────
Unexpected  49 9.37 5.41 |   8.67 1.64 23.29     0.68    -0.32
Expected    49 8.12 5.30 |   7.37 1.66 25.16     0.94     0.71
──────────────────────────────────────────────────────────────

Type III Analysis of Variance Table with Satterthwaite's method:
──────────────────────────────────────────────────────
         Sum Sq Mean Sq NumDF  DenDF     F     p
──────────────────────────────────────────────────────
Outcome  38.381  38.381 1.000 48.000 2.514  .119
──────────────────────────────────────────────────────
```

The analysis results are consistent with the original text: the *Inefficient Agent Task* ($F(1, 48) = 2.51$, $p =. 119$).

The *Instrumental-Action Task*

```
Exp2_instrumental_action<- read.csv('data/Exp2/Data from all participating
infants/instrumental_action.csv',
                                    encoding = 'UTF-8',
                                    header = TRUE) %>%
  P_cleaning() #读取并清理数据
P_describe(Exp2_instrumental_action)
```



```
model10<- lmer(Looking ~ Outcome + (1|Participant), data = Exp2_instrumental_action) #建立HLM模型
result10<- anova(model10) #对建立的模型进行 Type 3 Test
P_summary(result10) #呈现最终的结果
```

```
Descriptive Statistics:
──────────────────────────────────────────────────────────────
             N Mean   SD | Median  Min   Max Skewness Kurtosis
──────────────────────────────────────────────────────────────
Unexpected  48 8.09 5.58 |   7.02 1.51 28.25     1.54     2.72
Expected    48 7.92 5.86 |   5.66 1.65 25.23     1.05     0.15
──────────────────────────────────────────────────────────────

Type III Analysis of Variance Table with Satterthwaite's method:
──────────────────────────────────────────────────────────
         Sum Sq Mean Sq NumDF  DenDF     F     p
──────────────────────────────────────────────────────────
Outcome   0.698   0.698 1.000 47.000 0.035  .853
──────────────────────────────────────────────────────────
```

The analysis results are consistent with the original text: the *Instrumental Action Task* (*F* (1, 47)=0.03, *p*=. 853).

Repeat discussion

The results of both experiments suggest that infants have a rational expectation of effective agent behavior that generalizes to new and changing environments. New issues that may be related to common-sense abstraction principles are also raised, such as having the agent bypass the barrier to reach the target object may strengthen infants' goal attributions in that task; changes in the location of the barrier between trials may affect infants' assessment of target object accessibility; and the presence of a new agent receives high attention from infants.

In summary, we provided an overview and replication of Experiment 1 and Experiment 2. The original authors screened and described the data more explicitly; thus, the overall replication was not too difficult, and the replication results were largely consistent with the original literature results. At the same time, we also found the two errors just pointed out in the reproduction process. The first point is that,

The first questionable point is that during the data analysis of the efficient agent task in Experiment 1, we defined the same function and built the same model as the authors, but got different *F*-values. By examining the authors' publicly available R code, we believe that here the authors have filled in the wrong F-value. The second problem

found is that Experiment 2 gives inconsistent results with the original text when analyzing all the tasks done by the infant together. By examining the code shared by the authors, we found that the authors used the data from the pre-experiment when reading the data, thus reporting the results of the pre-experiment. When we used the data from the pre-experiment for our analysis, we got the same results. And in our analysis above, we used the data from the formal experiment, so we have reason to believe that the data we reported is the reasonable result of the analysis.

In the replication process, we not only gained a richer understanding of the study design and model validation, but also were able to find that the researchers still had omissions in some details, and thus an open, transparent, and open research orientation advocates that we can better promote the issue of reproducibility crisis in psychology. We hope that more researchers will join in the discussion of reproducibility studies to make these infinitely dynamic studies fuller and more rigorous.

General discussion

This literature examines the commonalities and differences between human infants and machines in understanding the expectations and intentions behind human behavior through a comparative BIB task framework.

BIB includes six highly minimal but presentationally consistent tasks focusing on three high-level principles of commonsense psychology: goal attribution; rationality attribution; and instrumentality attribution.

Infants' successes on BIB suggest they have a highly abstract notion of agents' actions as goal-directed towards objects and a principle of rationality that leads to default expectations of agents' efficient actions towards goals. These results are consistent with the rich literature on infants' commonsense psychology and synthesize the literature's findings in a unified framework that can be directly compared with - and perhaps built into - machine intelligence.

Infants' failures on BIB suggest that changes to the contexts in which goals are first demonstrated may have significant impacts on infants' goal and rationality attribution.

For example, infants may not generalize an agent's goal to a test environment with even minimal or inconsequential changes relative to the environment in which the goal was initially demonstrated if those changes suggest that agents are acting in a new place.

Further research has marked that infants and machines share some basic intuitions in understanding human behavior, but the latest learning-driven neural network models do not yet reach the level of common knowledge of infants on the existing tasks of BIB.

Nonetheless, extending and investigating the comprehensive framework for representing infant knowledge, e.g., expectations of agent cost and value concepts, or the recognition of agent behaviors as embodying social partnerships, has important implications for advances in artificial intelligence.

The division of work in this document:

1.Shu Kaiyan: document formatting, script layout, script synthesis, article introduction and results

2.Zhou Yingjie: article methods, discussion and conclusion, repeat idea and discussion

3.Xu Peng: repeat procedure, code review