

analysis for lai et al.

2023-12-22

In this document, we run the reliance framework analysis for deception detection task and its corresponding explanations in Lai et al. We estimate the behavioral agents' joint behavioral $\pi(\theta, v, a^b)$ by the empirical distribution in the experiment data. We first show the results using approximation of rational benchmark and the mis-reliant rational benchmark with overfitting to the empirical distribution and then show the results using approximation with the discretized signals.

```
con <- file('./label.json', "r")
raw_data <- ldply(fromJSON(con), data.frame)
raw_data = raw_data %>% group_by(review_text) %>% mutate(review_id = cur_group_id())
human_predictions = raw_data %>% filter(experiment == "control" | experiment == "highlight" |
                                         experiment == "heatmap" | experiment == "examples")
exp_data = raw_data %>% filter(experiment != "control" & experiment != "highlight" &
                               experiment != "heatmap" & experiment != "examples")
human_predictions = human_predictions %>%
  select(-predicted_label) %>%
  left_join(exp_data %>%
            select(predicted_label, review_text) %>%
            distinct(), by = c("review_text")) %>%
  mutate(predicted_label = as.numeric(predicted_label))

scoring_rule = function(action, state) {
  (action == state)
}

task_data = raw_data %>% group_by(review_text) %>% summarise(review_id = cur_group_id())
corpus = tm::Corpus(tm::VectorSource(task_data$review_text))
corpus.cleaned <- tm::tm_map(corpus, tm::removeWords, tm::stopwords('english')) # Removing stop-words

## Warning in tm_map.SimpleCorpus(corpus, tm::removeWords,
## tm::stopwords("english")): transformation drops documents
corpus.cleaned <- tm::tm_map(corpus, tm::stemDocument, language = "english") # Stemming the words

## Warning in tm_map.SimpleCorpus(corpus, tm::stemDocument, language = "english"):
## transformation drops documents
corpus.cleaned <- tm::tm_map(corpus.cleaned, tm::stripWhitespace) # Trimming excessive whitespaces

## Warning in tm_map.SimpleCorpus(corpus.cleaned, tm::stripWhitespace):
## transformation drops documents

tdm <- tm::DocumentTermMatrix(corpus.cleaned)
tdm.tfidf <- tm::weightTfIdf(tdm)
tdm.tfidf <- tm::removeSparseTerms(tdm.tfidf, 0.99)
tfidf.matrix <- as.matrix(tdm.tfidf)

predict.kmeans <- function(object, newdata){
  centers <- object$centers
```

```

n_centers <- nrow(centers)
dist_mat <- as.matrix(dist(rbind(centers, newdata)))
dist_mat <- dist_mat[-seq(n_centers), seq(n_centers)]
list(cluster = max.col(-dist_mat), total_error = sum(apply(dist_mat, 1, function(x) min(x))))
}

number_of_partition = 10
partition_size = nrow(tfidf.matrix) / number_of_partition
best.K = -1
best.benchmark = c()
best.test_sd = Inf
for (K in seq(10, 200, 10)) {
  benchmark = c()
  for (i in seq(1, nrow(tfidf.matrix), partition_size)) {
    test_set = tfidf.matrix[i:min(i + partition_size - 1, nrow(tfidf.matrix)),]
    training_set = tfidf.matrix[-(i:min(i + partition_size - 1, nrow(tfidf.matrix))),]
    clustering.kmeans <- kmeans(training_set, K)
    cluster_number = predict(clustering.kmeans, test_set)$cluster
    test_questionId = task_data$review_id[i:min(i + partition_size - 1, nrow(tfidf.matrix))]

    test_task_data = task_data %>% filter(review_id %in% test_questionId) %>% mutate(cluster = cluster_number)
    test_human_predictions = human_predictions %>%
      filter(review_id %in% test_questionId) %>%
      left_join(test_task_data %>% select(review_id, cluster), by = c("review_id"))

    train_task_data = task_data %>% filter(!(review_id %in% test_questionId)) %>% mutate(cluster = cluster_number)
    train_human_predictions = human_predictions %>%
      filter(!(review_id %in% test_questionId)) %>%
      left_join(train_task_data %>% select(review_id, cluster), by = c("review_id"))

    test_rational_action = train_human_predictions %>%
      rbind(test_human_predictions) %>%
      group_by(user_label, predicted_label, cluster) %>%
      mutate(pos_human_payoff = scoring_rule(user_label, actual_label),
             pos_ai_payoff = scoring_rule(predicted_label, actual_label)) %>%
      summarise(pos_human_payoff = mean(pos_human_payoff),
                pos_ai_payoff = mean(pos_ai_payoff))
    benchmark = c(benchmark, (test_rational_action %>%
      ungroup() %>%
      left_join(test_rational_action, by = c("user_label", "predicted_label", "cluster")) %>%
      mutate(benchmark_action = ifelse(pos_human_payoff > pos_ai_payoff, user_label, predicted_label)) %>%
      mutate(benchmark = scoring_rule(benchmark_action, actual_label)) %>%
      summarise(benchmark = mean(benchmark)))$benchmark)
  }
  if (best.test_sd > sd(benchmark)) {
    best.test_sd = sd(benchmark)
    best.benchmark = benchmark
    best.K = K
  }
}

```

```

## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.
## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can

```



```
## override using the `.groups` argument.
## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.
## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.
## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.
## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.
## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.
## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.
## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.
## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.
```

```
best.K
```

```
## [1] 50
```

```
clustering.kmeans <- kmeans(tfidf.matrix, best.K)
```

```
task_data = task_data %>% mutate(cluster = clustering.kmeans$cluster)
```

```
raw_data = raw_data %>% left_join(task_data, by = c("review_text", "review_id"))
```

```
human_predictions = raw_data %>% filter(experiment == "control" | experiment == "highlight" |
                                         experiment == "heatmap" | experiment == "examples")
```

```
exp_data = raw_data %>% filter(experiment != "control" & experiment != "highlight" &
                               experiment != "heatmap" & experiment != "examples")
```

```
human_predictions = human_predictions %>%
```

```
  select(-predicted_label) %>%
```

```
  left_join(exp_data %>%
```

```
    select(predicted_label, review_text) %>%
```

```
    distinct(), by = c("review_text")) %>%
```

```
  mutate(predicted_label = as.numeric(predicted_label))
```

```
exp_data = exp_data %>%
```

```
  left_join(human_predictions %>%
```

```
    select(human_pred = user_label, review_text),
```

```
    by = c("review_text")) %>%
```

```
  mutate(predicted_label = as.numeric(predicted_label))
```

```
## Warning in left_join(., human_predictions %>% select(human_pred = user_label, : Detected an unexpected
```

```
## i Row 1 of `x` matches multiple rows in `y`.
```

```
## i Row 226 of `y` matches multiple rows in `x`.
```

```
## i If a many-to-many relationship is expected, set `relationship =
```

```
## "many-to-many" to silence this warning.
```

Approximating by overfitting to the empirical distribution

```
rational_action = human_predictions %>%
```

```
  group_by(user_label, predicted_label, review_text) %>%
```

```
  mutate(pos_human_payoff = scoring_rule(user_label, actual_label),
```

```

    pos_ai_payoff = scoring_rule(predicted_label, actual_label)) %>%
  summarise(pos_human_payoff = mean(pos_human_payoff),
            pos_ai_payoff = mean(pos_ai_payoff))

## `summarise()` has grouped output by 'user_label', 'predicted_label'. You can
## override using the `.groups` argument.

rational_data = human_predictions %>%
  ungroup() %>%
  mutate(prior_human_payoff = mean(scoring_rule(user_label, actual_label)),
         prior_ai_payoff = mean(scoring_rule(predicted_label, actual_label))) %>%
  mutate(baseline_action = ifelse(prior_human_payoff > prior_ai_payoff, user_label, predicted_label)) %>%
  mutate(baseline = scoring_rule(baseline_action, actual_label)) %>%
  mutate(baseline2_action = ifelse(prior_human_payoff <= prior_ai_payoff, user_label, predicted_label)) %>%
  mutate(baseline2 = scoring_rule(baseline2_action, actual_label)) %>%
  left_join(rational_action) %>%
  mutate(benchmark_action = ifelse(pos_human_payoff > pos_ai_payoff, user_label, predicted_label)) %>%
  mutate(benchmark = scoring_rule(benchmark_action, actual_label)) %>%
  mutate(rational_reliance_level = (benchmark_action == predicted_label) & (user_label != predicted_label))

## Joining with `by = join_by(user_label, review_text, predicted_label)`
rational_data

## # A tibble: 6,400 x 24
##   anonymized_worker_id experiment total_time_in_minutes user_label actual_label
##   <chr>                  <chr>                <dbl>      <dbl>      <dbl>
## 1 5AB84043BCE99         control                10          0          0
## 2 5AB84043BCE99         control                10          1          0
## 3 5AB84043BCE99         control                10          1          0
## 4 5AB84043BCE99         control                10          1          0
## 5 5AB84043BCE99         control                10          1          0
## 6 5AB84043BCE99         control                10          1          1
## 7 5AB84043BCE99         control                10          0          1
## 8 5AB84043BCE99         control                10          0          1
## 9 5AB84043BCE99         control                10          0          1
## 10 5AB84043BCE99        control                10          0          1
## # i 6,390 more rows
## # i 19 more variables: review_time_in_seconds <chr>, review_num <dbl>,
## #   deceptive <dbl>, positive <dbl>, review_text <chr>, review_id <int>,
## #   cluster <int>, predicted_label <dbl>, prior_human_payoff <dbl>,
## #   prior_ai_payoff <dbl>, baseline_action <dbl>, baseline <lgl>,
## #   baseline2_action <dbl>, baseline2 <lgl>, pos_human_payoff <dbl>,
## #   pos_ai_payoff <dbl>, benchmark_action <dbl>, benchmark <lgl>, ...

sample_size = 40
n_round = 1000
rational_results = data.frame()
for (i in 1:n_round) {
  rational_results = rational_data %>%
    group_by(review_num) %>%
    sample_n(sample_size) %>%
    ungroup() %>%
    summarise(baseline = mean(baseline),
              baseline2 = mean(baseline2),
              benchmark = mean(benchmark),

```

```

        rational_reliance = mean(rational_reliance_level)) %>%
    rbind(rational_results)
}
rational_results = rational_results %>%
    cross_join(exp_data %>% group_by(experiment) %>% summarise())
rational_results

## # A tibble: 5,000 x 5
##   baseline baseline2 benchmark rational_reliance experiment
##   <dbl>      <dbl>      <dbl>          <dbl> <chr>
## 1  0.878      0.556      0.942          0.386 machine
## 2  0.878      0.556      0.942          0.386 machine_and_examples
## 3  0.878      0.556      0.942          0.386 machine_and_heatmap
## 4  0.878      0.556      0.942          0.386 machine_and_random_heatmap
## 5  0.878      0.556      0.942          0.386 machine_with_accuracy
## 6  0.859      0.552      0.934          0.381 machine
## 7  0.859      0.552      0.934          0.381 machine_and_examples
## 8  0.859      0.552      0.934          0.381 machine_and_heatmap
## 9  0.859      0.552      0.934          0.381 machine_and_random_heatmap
## 10 0.859      0.552      0.934          0.381 machine_with_accuracy
## # i 4,990 more rows

behavioral_data = exp_data %>%
    mutate(behavioral = scoring_rule(user_label, actual_label))

sample_size = 40
n_round = 1000
behavioral_results = data.frame()
for (i in 1:n_round) {
    behavioral_result = behavioral_data %>%
        group_by(review_num, experiment) %>%
        sample_n(sample_size) %>%
        group_by(experiment) %>%
        summarise(behavioral = mean(behavioral))
    reliance_level = behavioral_data %>%
        group_by(review_num, experiment) %>%
        sample_n(sample_size) %>%
        mutate(sample_id = row_number()) %>%
        group_by(experiment, sample_id) %>%
        summarise(reliance_level = mean((user_label == predicted_label) & (human_pred != user_label))) %>%
        group_by(experiment) %>%
        summarise(reliance_level = mean(reliance_level))
    misreliant = rational_data %>%
        group_by(review_num) %>%
        sample_n(sample_size) %>%
        mutate(sample_id = row_number()) %>%
        select(-experiment) %>%
        cross_join(reliance_level) %>%
        group_by(sample_id, experiment) %>%
        arrange(desc(pos_ai_payoff - pos_human_payoff), .by_group = TRUE) %>%
        mutate(sort_id = row_number()) %>%
        mutate(max_sort_id = max(sort_id)) %>%
        mutate(misreliant_action = ifelse(sort_id <= reliance_level * max_sort_id,
            predicted_label,

```

}

1111 `summarise()` has grouped output by `experiment`. You can override using `uns`

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
```

```
behavioral_results
```

```
## # A tibble: 5,000 x 4
##   experiment      behavioral misreliant reliance_level
##   <chr>          <dbl>      <dbl>      <dbl>
## 1 machine          0.615      0.816      0.261
## 2 machine_and_examples 0.694      0.889      0.371
## 3 machine_and_heatmap  0.702      0.889      0.37
## 4 machine_and_random_heatmap 0.679      0.889      0.356
## 5 machine_with_accuracy 0.775      0.889      0.354
## 6 machine          0.634      0.831      0.308
## 7 machine_and_examples 0.69      0.869      0.351
## 8 machine_and_heatmap  0.722      0.869      0.356
## 9 machine_and_random_heatmap 0.716      0.869      0.365
## 10 machine_with_accuracy 0.748      0.869      0.364
## # i 4,990 more rows
```

```
results = rational_results %>%
  group_by(experiment) %>%
  summarise(benchmark = mean(benchmark), baseline = mean(baseline))
results = results %>%
  left_join(behavioral_results %>%
    group_by(experiment) %>%
    summarise(behavioral = mean(behavioral),
              misreliant = mean(misreliant)))
```

```
## Joining with `by = join_by(experiment)`
```

```
results = results %>%
  mutate(belief_loss = (misreliant - behavioral)/(benchmark - baseline),
         reliance_loss = (benchmark - misreliant)/(benchmark - baseline)) %>%
  arrange(reliance_loss)
results
```

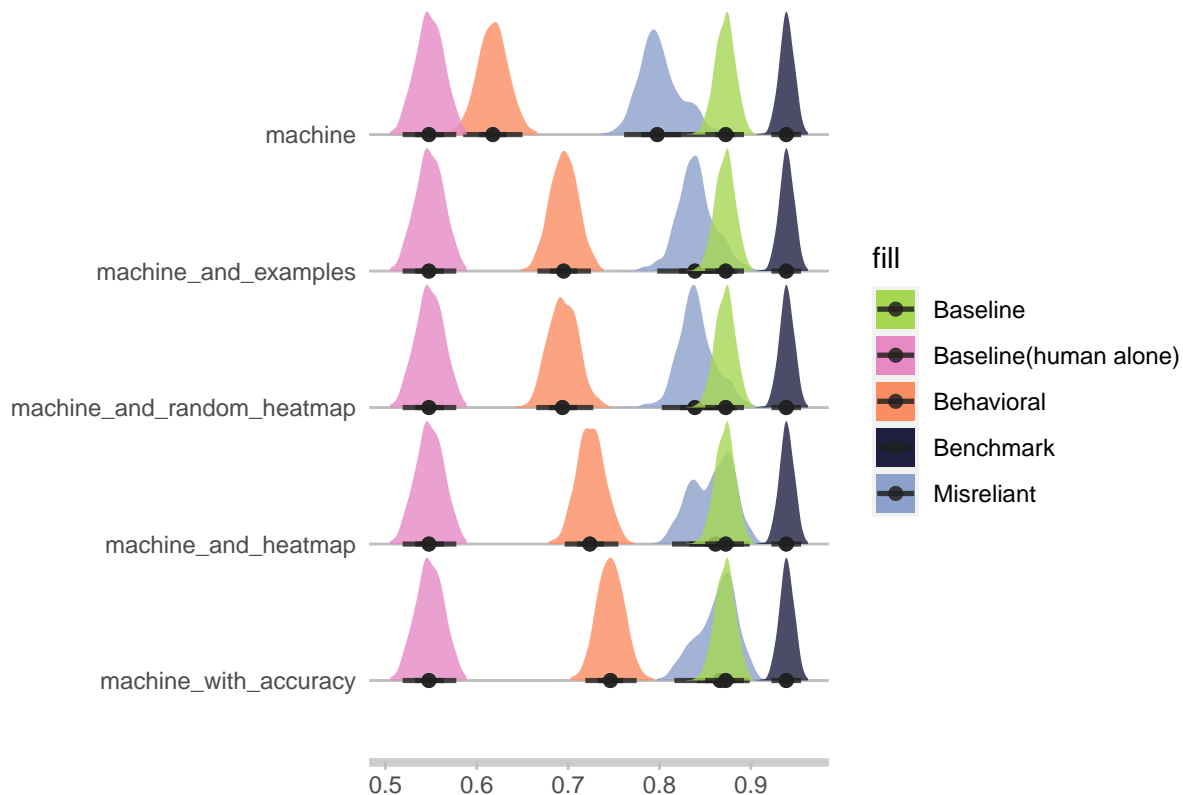
```
## # A tibble: 5 x 7
##   experiment benchmark baseline behavioral misreliant belief_loss reliance_loss
##   <chr>          <dbl>    <dbl>    <dbl>      <dbl>      <dbl>      <dbl>
## 1 machine_wi~  0.939    0.872    0.746    0.863      1.73      1.13
## 2 machine_an~  0.939    0.872    0.725    0.858      1.97      1.21
## 3 machine_an~  0.939    0.872    0.695    0.841      2.16      1.46
## 4 machine_an~  0.939    0.872    0.695    0.840      2.14      1.47
## 5 machine      0.939    0.872    0.618    0.801      2.72      2.05
```

```
colors <- c("Baseline" = "#a6d854", "Baseline(human alone)" = "#e78ac3", "Benchmark" = "#1f2041", "Behavioral")
ggplot() +
```

```

stat_slabinterval(data = behavioral_results, aes(y = experiment, x = behavioral, fill = "Behavioral"))
stat_slabinterval(data = behavioral_results, aes(y = experiment, x = misreliant, fill = "Misreliant"))
stat_slabinterval(data = rational_results, aes(y = experiment, x = baseline, fill = "Baseline"), alpha = 0.5)
stat_slabinterval(data = rational_results, aes(y = experiment, x = baseline2, fill = "Baseline(human alone)"), alpha = 0.5)
stat_slabinterval(data = rational_results, aes(y = experiment, x = benchmark, fill = "Benchmark"), alpha = 0.5)
labs(x = "", y = "", color = "Quantity") +
ylim((results)$experiment) +
theme(panel.grid.major.x = element_blank(),
      panel.grid.minor.x = element_blank(),
      panel.grid.major = element_line(colour = "grey"),
      axis.line.x = element_line(linewidth = 1.5, colour = "grey80"),
      panel.background = element_rect(fill = "white", color = "white"),
      axis.ticks.y = element_blank(),
      axis.ticks.x = element_line(colour = "grey")) +
scale_fill_manual(values = colors)

```



```

# ggsave("./lai_results.pdf", unit = "in", width = 7.22222222222, height = 3.47222222222 * 5/4)

```

Using discretized signals to approximate

```

rational_action = human_predictions %>%
  group_by(user_label, predicted_label, cluster) %>%
  mutate(pos_human_payoff = scoring_rule(user_label, actual_label),
         pos_ai_payoff = scoring_rule(predicted_label, actual_label)) %>%
  summarise(pos_human_payoff = mean(pos_human_payoff),
            pos_ai_payoff = mean(pos_ai_payoff))

```

`summarise()` has grouped output by 'user_label', 'predicted_label'. You can
override using the `.groups` argument.

```
rational_data = human_predictions %>%
  ungroup() %>%
  mutate(prior_human_payoff = mean(scoring_rule(user_label, actual_label)),
         prior_ai_payoff = mean(scoring_rule(predicted_label, actual_label))) %>%
  mutate(baseline_action = ifelse(prior_human_payoff > prior_ai_payoff, user_label, predicted_label)) %>%
  mutate(baseline = scoring_rule(baseline_action, actual_label)) %>%
  mutate(baseline2_action = ifelse(prior_human_payoff <= prior_ai_payoff, user_label, predicted_label)) %>%
  mutate(baseline2 = scoring_rule(baseline2_action, actual_label)) %>%
  left_join(rational_action) %>%
  mutate(benchmark_action = ifelse(pos_human_payoff > pos_ai_payoff, user_label, predicted_label)) %>%
  mutate(benchmark = scoring_rule(benchmark_action, actual_label)) %>%
  mutate(rational_reliance_level = (benchmark_action == predicted_label) & (user_label != predicted_label))
```

Joining with `by = join_by(user_label, cluster, predicted_label)`

```
sample_size = 40
n_round = 1000
rational_results = data.frame()
for (i in 1:n_round) {
  rational_results = rational_data %>%
    group_by(review_num) %>%
    sample_n(sample_size) %>%
    ungroup() %>%
    summarise(baseline = mean(baseline),
              baseline2 = mean(baseline2),
              benchmark = mean(benchmark),
              rational_reliance = mean(rational_reliance_level)) %>%
    rbind(rational_results)
}
rational_results = rational_results %>%
  cross_join(exp_data %>% group_by(experiment) %>% summarise())

behavioral_data = exp_data %>%
  mutate(behavioral = scoring_rule(user_label, actual_label))

sample_size = 40
n_round = 1000
behavioral_results = data.frame()
for (i in 1:n_round) {
  behavioral_result = behavioral_data %>%
    group_by(review_num, experiment) %>%
    sample_n(sample_size) %>%
    group_by(experiment) %>%
    summarise(behavioral = mean(behavioral))
  reliance_level = behavioral_data %>%
    group_by(review_num, experiment) %>%
    sample_n(sample_size) %>%
    mutate(sample_id = row_number()) %>%
    group_by(experiment, sample_id) %>%
    summarise(reliance_level = mean((user_label == predicted_label) & (human_pred != user_label))) %>%
    group_by(experiment) %>%
    summarise(reliance_level = mean(reliance_level))
}
```

```

misreliant = rational_data %>%
  group_by(review_num) %>%
  sample_n(sample_size) %>%
  mutate(sample_id = row_number()) %>%
  select(-experiment) %>%
  cross_join(reliance_level) %>%
  group_by(sample_id, experiment) %>%
  arrange(desc(pos_ai_payoff - pos_human_payoff), .by_group = TRUE) %>%
  mutate(sort_id = row_number()) %>%
  mutate(max_sort_id = max(sort_id)) %>%
  mutate(misreliant_action = ifelse(sort_id <= reliance_level * max_sort_id,
                                     predicted_label,
                                     user_label)) %>%

  mutate(misreliant = scoring_rule(misreliant_action, actual_label)) %>%
  group_by(experiment) %>%
  summarise(misreliant = mean(misreliant), reliance_level = mean(reliance_level))

behavioral_results = behavioral_result %>%
  left_join(misreliant, by = c("experiment")) %>%
  rbind(behavioral_results)
}

```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

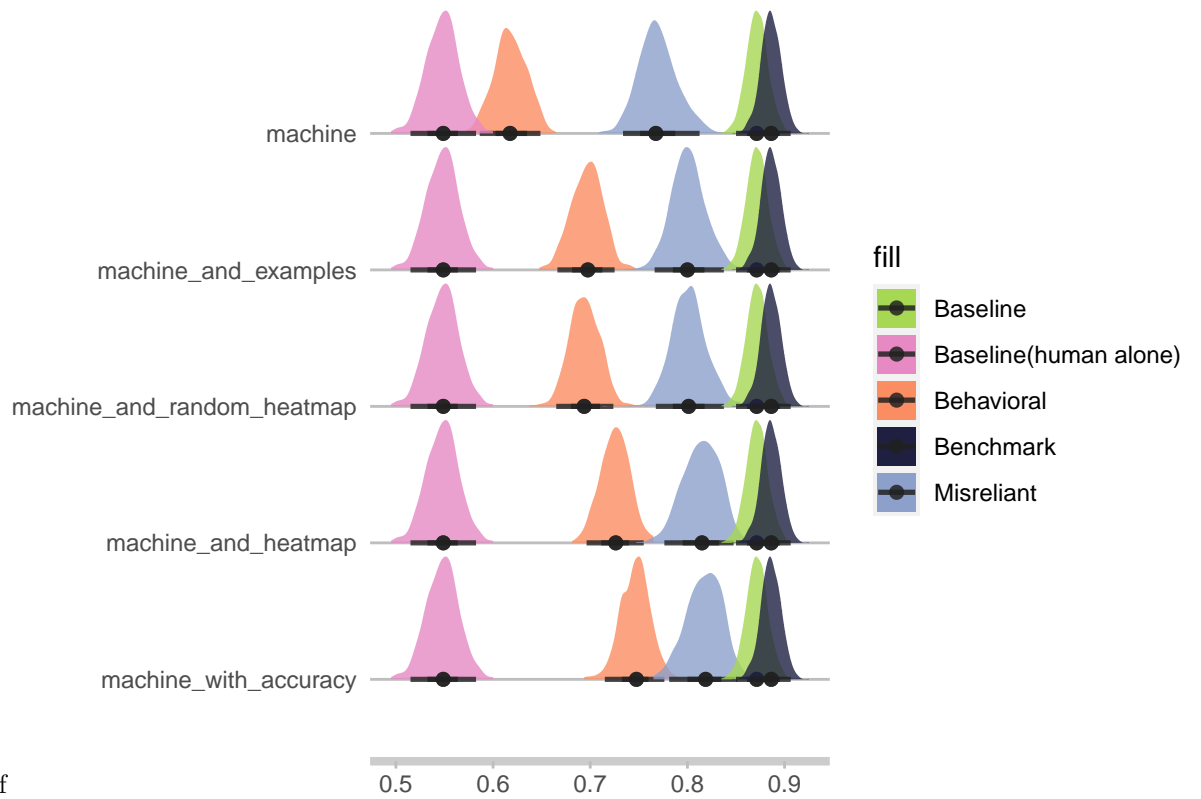
[illegible]


```
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'experiment'. You can override using the
## `.groups` argument.
```

```
results = rational_results %>%
  group_by(experiment) %>%
  summarise(benchmark = mean(benchmark), baseline = mean(baseline))
results = results %>%
  left_join(behavioral_results %>%
    group_by(experiment) %>%
    summarise(behavioral = mean(behavioral),
              misreliant = mean(misreliant)))
```

```
## Joining with `by = join_by(experiment)`
```

```
results = results %>%
  mutate(belief_loss = (misreliant - behavioral)/(benchmark - baseline),
         reliance_loss = (benchmark - misreliant)/(benchmark - baseline)) %>%
  arrange(reliance_loss)
ggplot() +
  stat_slabinterval(data = behavioral_results, aes(y = experiment, x = behavioral, fill = "Behavioral"))
  stat_slabinterval(data = behavioral_results, aes(y = experiment, x = misreliant, fill = "Misreliant"))
  stat_slabinterval(data = rational_results, aes(y = experiment, x = baseline, fill = "Baseline"), alpha = 0.5)
  stat_slabinterval(data = rational_results, aes(y = experiment, x = baseline2, fill = "Baseline(human)"), alpha = 0.5)
  stat_slabinterval(data = rational_results, aes(y = experiment, x = benchmark, fill = "Benchmark"), alpha = 0.5)
  labs(x = "", y = "", color = "Quantity") +
  ylim((results)$experiment) +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.major = element_line(colour = "grey"),
        axis.line.x = element_line(linewidth = 1.5, colour = "grey80"),
        panel.background = element_rect(fill = "white", color = "white"),
        axis.ticks.y = element_blank(),
        axis.ticks.x = element_line(colour = "grey")) +
  scale_fill_manual(values = colors)
```



signals-1.pdf

```
# ggsave("./lai_results_test_performance.pdf", unit = "in", width = 7.2222222222, height = 3.4722222222)
```