Backward Induction and Empirical Complexity: A Mobile Experiment

Konrad Grabiszewski*    Alex Horenstein †

February 2018

Abstract

We experimentally analyze violations of backward induction using 27 trees. We rank trees by empirical complexity; that is, the percentage of subjects who do not behave in accordance with backward induction. As typical in the literature, we find that, in every tree, empirical complexity is more than 0%. However, given the richness of our data, we find more: empirical complexity is not the same in each tree but in fact ranges from 2% to 49%. In order to explain the variation in empirical complexity, we need to look beyond the typical explanations of deviations from backward induction that are found in the literature: it is necessary to look at the variation in tree structure. We find that expanding tree length (more rounds), rather than tree width (more actions), is the most important factor determining empirical complexity. Our data comes from an iOS/Android mobile game Blues and Reds that we developed for smartphones and tablets.

Keywords: game theory; finite tree; backward induction; empirical complexity; mobile experiment

JEL classification: C72, C73, C99

*Department of Economics, University of Miami, Coral Gables, FL 33124, USA; konrad.grabiszewski@gmail.com
†Department of Economics, University of Miami, Coral Gables, FL 33124, USA; alexhorenstein@gmail.com
1 Introduction

It can be said that the success of a scientific concept is measured by the number of studies that attempt to dismiss it. From that perspective, backward induction is immensely successful. Although fundamental in theoretical and applied economics, thousand of experimental studies are less than enthusiastic about backward induction. The general consensus is that people do not behave as if following backward induction reasoning.

The literature offers many explanations for the discrepancy between the behavior that we observe and the behavior that backward induction predicts. We add to this literature and look at the problem of backward induction from a perspective that, so far, has been neglected: the role of the tree structure.

We focus on the tree structure because the failure rate of backward induction — that is, the percentage of subjects who violate backward induction — is not the same across different dynamic games. While subject-based explanations found in the literature (e.g., limited abilities) tell us why the failure rates differ from zero, these explanation can not tell us why the failure rates are not identical. To understand what drives the differences in the failure rates, it is necessary to look at the differences in the games (i.e., tree structure) that subjects participate in. This calls for an experiment with many trees.

Our data comes from a mobile experiment: in order to gather data, we created a mobile game Blues and Reds that has been available for free for iOS and Android devices. We use the data from 27 turn-based, finite dynamic games with perfect and complete information. In each game, subjects play against Artificial Intelligence (AI). In each game, there are only two possible outcomes: subject wins (AI loses) or subject loses (AI wins). Our project achieves two objectives.

Objective 1: Which tree is more complex? First, we let the data speak and order trees from the least to the most complex. Our measure of empirical complexity is the percentage of subjects whose behavior in a tree is not consistent with backward induction. Higher percentage means that a tree is empirically more complex.

In some cases, comparing trees requires no empirical work. One tree can be objectively less complex than another tree. However, it is not possible to rank all trees by their complexity without data.
To elaborate, consider the three trees depicted below: A (Figure 1), B (Figure 2), and C (Figure 3) with payoffs in order \((Ann, AI)\).

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

Tree A is objectively less complex than both trees B and C. Hence, the percentage of subjects who choose \(R\) in tree A should not be bigger than the percentage of subjects playing \(L\) or \(R\) in tree B and the percentage of subjects not playing \(R, L_3\) or \(R, L_4\) in tree C. However, it is not clear which tree – B or C – is less complex. Figuring that out empirically is precisely the first objective of our paper.

**Objective 2: What drives complexity?** Our second objective is to analyze the variation in empirical complexity across the trees. That complexity ranges from 2\% to 49\%. Why do people behave in accordance with backward induction in some trees but violate backward induction in other trees? The richness of our data allows us to address this question. In particular, we investigate whether it is the length or width of the tree that has the most significant impact on empirical complexity.

When we start with a given tree, then we can add more rounds (make the tree longer) or more actions at the already existing non-final nodes (make the tree wider). For example, tree B is constructed from tree A by adding one more action at each non-final node, and tree C is an extension of tree A by adding one more round.

It turns out that it is the length rather than the width that makes the tree more complex. In short, people struggle more with longer time horizons rather than larger sets of actions at each round.

Our results are important for several reasons. First, they enrich our understanding of people’s cognitive limits. From the behavioral literature, we already know that people have limited capabilities when it comes to backward inducting. Our results go further and show not only that limited capabilities affect behavior but also how they affect behavior in the context of tree structure.
Second, our results are of interest from the mechanism design perspective. Consider a regulator who designs a new law (dynamic game) in order to achieve specific outcome. Suppose that the regulator has two design choices: a game with fewer options per round but many rounds, and a game with more options per round but fewer rounds. While under the assumption of unlimited skills, both games are equivalent, but our results indicate that in a more realistic setup with limited skills, it is better to choose a shorter game.

Our paper is constructed in the following way. In Section 2, we review the literature. In Section 3, we discuss the design of our experiment. Section 4 contains our empirical analysis. Section 5 closes the paper.

2 Literature Review

We divide the literature review into two parts in accordance with the two objectives described in the Introduction.

2.1 Objective 1: Which tree is more complex?

To the best of our knowledge, there is no paper that uses a multiplicity of trees in an attempt to rank these trees according to their empirical measure of complexity. If a study experiments with more than one tree, then, on purposes, the trees can be objectively ranked by their complexity.

McKelvey and Palfrey (1992) and Fey et al. (1996) experiment with the centipede game of different lengths; 4-move and 6-move in the former, and 6-move and 10-move in the latter. Not surprisingly, both find that backward induction is more likely to be violated in a longer centipede. Dufwenberg et al. (2010) and Gneezy et al. (2010) use two forms of the race game; one larger than the other. Each finds that the failure rate of backward induction is higher in a larger game.

2.2 Objective 2: What drives complexity?

Why people do not behave in accordance with backward induction has been extensively studied in the literature. The empirical literature on backward induction can be divided into two streams:
limited abilities and alternative preferences. Our study belongs to the first stream. However, as far as we know, ours is the first analysis that links tree structure to subjects’ limited abilities to backward induct.

### 2.2.1 Limited abilities

People behave as if violating backward induction due to their limited abilities (a) to solve the game or (b) to understand the game they play.

(a) *Limited ability to solve the game.* In the literature, it has been established that a player’s cognitive skills not only correlate with her strategic behavior being more consistent with the theory\(^1\) but also influence the behavior of other players who take into account limited abilities of their opponents.\(^2\)

The fundamental model of imperfect strategic reasoning is the level-\(k\) model.\(^3\) In this model, players are partitioned into types that indicate their reasoning skills. Level-0 players are non-strategic, level-1 best-respond to level-0 players, etc. This model has been extensively tested in the literature and the main message is that people, indeed, struggle with strategic reasoning.\(^4\)

(b) *Limited ability to understand the game.* Before players can employ their skills to solve the game, it is necessary for them to understand the game they play. The literature supports the view that people struggle with game form recognition.

In contradiction to theory, people behave differently when presented with two equivalent but different forms of the same game.\(^5\) Subjects do not properly see the relationship between choices

---

\(^1\)E.g., Burks et al. (2009), Burnham et al. (2009), Rydval et al. (2009), Brañas-Garza et al. (2012), Carpenter et al. (2013), Duffy and Smith (2014), Agranov et al. (2015), Allred et al. (2016), Bayer and Renou (2016b), Benito-Ostolaza et al. (2016), Gill and Prowse (2016), Hanaki et al. (2016), and Kiss et al. (2016). See also Rubinstein (2007) and Rubinstein (2016) which use response time to identify instinctive and contemplative (cognitive) types.

\(^2\)E.g., Palacios-Huerta and Volij (2009), Agranov et al. (2012), Alaoui and Penta (2016), Fehr and Huck (2016), and Gill and Prowse (2016).

\(^3\)This model was introduced in Stahl and Wilson (1994), Stahl and Wilson (1995), and Nagel (1995). See also the closely related model of cognitive hierarchy developed in Camerer et al. (2004).


\(^5\)E.g., Schotter et al. (1994), Rapoport (1997), McCabe et al. (2000), and Cox and James (2012).
and outcomes. In fact, people do not even search for relevant information about the problem they are trying to solve. Confusion due to game-form has been used to rationalize, for example, cooperation observed in the public-good experiments. In our companion paper Grabiszewski and Horenstein (2017), we also address the issue of inability to correctly recognize the game-form.

### 2.2.2 Alternative preferences

When it comes to testing for backward induction, and nothing but backward induction, it is necessary to experiment on games that are not affected by other phenomena. For example, the race game and the games we use in this paper allow for unambiguous testing of backward induction.

However, sometimes we would like to test a joint hypothesis: subjects exhibit alternative preferences (e.g., fairness, altruism) and subjects backward induct. This would be the case of experiments with the very popular ultimatum game and centipede game. Since these experiments test for more than just backward induction, they differ from the “limited abilities” literature and our paper. The behavior observed in these experiments should not be interpreted as a rejection of backward induction; rather, these experiments do not reject the joint hypothesis they test.

In the ultimatum game (Rosenthal (1981)), which is the simplest version of the bargaining model (Stahl (1972) and Rubinstein (1982)), Ann and Bob have $100 to be divided among them. Ann makes an offer \((100 - x, x)\) that Bob either accepts or reject. If he accepts the offer, then Ann gets \(100 - x\) and \(x\) is for Bob. If Bob rejects the offer, then each player gets $0.

If we assume that a player’s utility function depends in an increasing fashion only on her/his monetary gain, then backward induction predicts that the first player (Ann) offers \(x\) as low as possible and the second player (Bob) accepts the offer. Not surprisingly, in the first experimental study of the ultimatum game (Güth et al. (1982)), and the thousands of others that have followed it, the observed results have contradicted this prediction. According to Güth and Kochner (2014), players in the role of Ann on average offer 40-50% of the pie and that offer is usually accepted.\(^9\)

---

\(^6\)E.g., Chou et al. (2009), Rydval et al. (2009), and Cason and Plott (2014).

\(^7\)Searching for information has been studied in the context of decision theory (e.g., Arieli et al. (2011) and Reutskaja et al. (2011)), normal-form games (e.g., Costa-Gomes et al. (2001), Salmon (2004), Costa-Gomes and Crawford (2006), Knoepflé et al. (2009), Brocas et al. (2014), and Devetag et al. (2016)), and extensive-form games (e.g., Johnson et al. (2002) and Wang et al. (2010)).

\(^8\)E.g., Andreoni (1995), Ferraro and Cummings (2007), and Houser and Kurzban (2002).

\(^9\)Note, however, that the value of the pie that Ann and Bob are to share also determines the behavior: when the pie gets bigger, smaller offers are accepted more often (Andersen et al. (2011)).
The most common explanation for what we observe in the ultimatum game is the notion of fairness that we can capture in the following way: Bob’s utility is constant if \( x \) is below a certain threshold (fair share), and only strictly increases above that threshold.

If Ann knows that Bob cares about fairness, then — even if she only cares about her own monetary gains and is not driven by any fairness concerns — she would make an offer that is at least equal to Bob’s fairness threshold. In other words, if Ann offers Bob a 45% share in the pie and Bob accepts it, then we should not conclude that backward induction is violated. Rather, we conclude that the joint hypothesis — people take fairness into account \( \text{and} \) people backward induct — is not rejected.\(^\text{10}\)

In the centipede game (Rosenthal (1981)), two players alternate in deciding whether to stop or go. If a player chooses to stop, then she/he gains more (in terms of monetary gains) compared to what she/he would gain if the opponent stops at the next round but less if the opponent chooses to go.

Under the standard assumptions, backward induction predicts that the first player stops the game at the first node. Since the first experiment involving the centipede game (McKelvey and Palfrey (1992)), the literature has repeatedly rejected this prediction. A typical explanation is the notion of altruism. As McKelvey and Palfrey (1992) indicate, “if it is believed that there is some likelihood that each player may be an altruist, then it can pay a selfish player to try to mimic the behavior of an altruist in an attempt to develop a reputation for passing.” In other words, the centipede game does not reject the presence of backward induction. Rather, this game tests, and does not reject, the joint hypothesis of altruism \( \text{and} \) backward induction.\(^\text{11}\)

\(^\text{10}\)Other rationalizations have also been provided: culture (Oosterbeek et al. (2004), Chuah et al. (2009), Chen and Tang (2009), Ferraro and Cummings (2007), and Boarini et al. (2009)), information about the personal features of players like name and physical description (Marchetti et al. (2011) and Charness and Gneezy (2008)), masculinity as measured by level of testosterone (Burnham (2007)), and pre-play communication (Zultan (2012) and Lusk and Hudson (2004)). See Güth and Kochner (2014) for the most recent survey of experiments based on the ultimatum game. As far as we know, Binmore et al. (2002) is the only example of unambiguous test of BI using a bargaining model. Following Harsanyi and Selten (1988), they split backward induction into two components, sub-game consistency and truncation consistency, which they test and reject.

\(^\text{11}\)Other explanations include mistakes (Fey et al. (1996)), payoff uncertainty (Zauner (1999)), lack of (common) knowledge of rationality (Palacios-Huerta and Volij (2009)), and non-standard preferences (Eichberger et al. (2017). See Krockow et al. (2016) for the most recent survey of experiments based on the centipede game.
3 Experimental Design

We use the data from a mobile game *Blues and Reds* that we developed for iOS and Android devices. We analyze behavior in 27 finite dynamic games with perfect and complete information. Each game is depicted as a tree; see an example in Figure 4.

In each tree, the subject plays against AI in turns with the subject always starting the game. There are two possible outcomes: the subject wins/AI loses and the subject loses/AI wins. This very simple structure allows us to disregard issues like payoff uncertainty (Zauner (1999)). Subjects play against AI rather than against each other because it removes the problems of social preferences (Johnson et al. (2002)). Winning is rewarded with in-game awards (stars and diamonds).

Each tree has four important properties.

1. Tree is winnable; i.e., subject can win.
2. In order to win, subject must not make a mistake at any round; i.e., making a mistake guarantees a loss.
3. At the initial node, there is only one action that would lead to the subject winning. This property holds at the subsequent nodes that lead towards a win.
4. The subject has only one opportunity to play a given tree, which, together with incentives provided in *Blues and Reds* (stars and diamond), motivates the subject to carefully analyze the problem and think when making her choices.

Given the properties of trees we use in the experiment, winning in a tree is equivalent to backward inducting. That is, if a subject wins, then we say that she behaves as if backward inducting.

In order to make conclusions related to the tree structure, all trees can be described as $N_1.N_2.N_3.N_4.N_5.N_6$ where $N_i$ denotes the number of actions at each node at the round $i$. In our sample, $N_i \in \{2, 3, 4\}$. To simplify notation, we write, for instance, 3.3 instead of 3.3.0.0.0.0. In Figure 4, there is tree 3.2.3.
We chose the tree structure $N_1.N_2.N_3.N_4.N_5.N_6$ because we want to explain what drives the empirical complexity in terms of tree structure. In particular, this requires that every path (from the initial node to a final node) has the same length.

To elaborate, consider tree 1 and tree 2 in Figures 5 and 6, respectively. In each tree, a quick glance at the payoffs indicates that Ann should play $\beta_1$ at her first node. There is no need to exert effort on further analysis. It does not matter that tree 2 is bigger than tree 1 — we expect the same percentage of wins and loses in each tree. Consequently, we learn nothing about the impact of tree structure on behavior.

[Figure 5 about here.]

[Figure 6 about here.]

Two trees — 4.2 and 2.4.2 — were used as a mandatory tutorial in which subjects were forced to learn about the rules and objectives of Blues and Reds. We do not use include these two trees in our data because every subject was guided how to win them.

The list of trees is presented in Table 1.

[Table 1 about here.]

While our goals are to rank trees according to their empirical complexity and understand what drives this ranking, in some cases the comparison between two trees can be done without any empirical studies. Some trees are objectively more complex than other trees. We say that tree $N_1.N_2.N_3.N_4.N_5.N_6$ is objectively more complex than tree $M_1.M_2.M_3.M_4.M_5.M_6$ if

1. $N_i \geq M_i$ for each $i$ with at least one inequality being strict, or

2. tree $M_1.M_2.M_3.M_4.M_5.M_6$ consists of $k$ rounds (i.e., $M_l = 0$ for $l > k$) and there exists $j$ such that $N_j \geq M_1$, $N_{j+1} \geq M_2$, ..., $N_{j+k} \geq M_k$ with at least one inequality being strict.

When tree $X$ is objectively more complex than tree $Y$, then we denote it by $X \succeq Y$. According to our definition of “objectively more complex,” 2.2.2 is objectively more complex than 2.2, and 2.4.2 is objectively more complex than 3.2. However, 2.3 is not objectively less/more complex than 3.2.2
nor is $3.3$ objectively more/less complex than $3.2.3$. Our main focus is to identify and explain the non-objective comparisons.

All objective pairwise comparisons of complexity among the trees in our sample are depicted in Figure 7. In that graph, the coordinates are the trees. We read that graph in the following way: take a tree $X$ from the vertical axis and a tree $Y$ from the horizontal axis. If the symbol at the intersection of $X$th row and $Y$th column is $\triangleright$, then $X \triangleright Y$, and if the symbol is a dot, then it is not possible to objectively compare $X$ and $Y$.

[Figure 7 about here.]

4 Empirical Analysis

4.1 Data characteristics

We collected data from August 15, 2017 to October 30, 2017. In that period, the game was downloaded over 20,000 times. In our analysis, we only use data from 2,097 subjects who played at least one tree of Blues and Reds beyond the mandatory tutorial. Our final sample includes subjects from at least 110 countries with the largest proportion coming from the USA (16%), followed by Mexico (13%), Argentina (10%), India (6%), Poland (5%), Thailand (5%), and Brazil (4%).

Table 2 shows the total number of subjects in each of the 27 trees.

[Table 2 about here.]

4.2 Empirical complexity: Ranking

Our first objective is to rank trees according to their empirical complexity, which we define as the percentage of subjects who violate backward induction (that is, they lose). Figure 8 shows the trees ranked in ascending order by their empirical complexity.

[Figure 8 about here.]
According to Figure 8, the least complex game is tree 2.4 (only 2.2% of the subjects lost in this tree), while the most complex one is tree 2.2.2.2.2 (49.3% of the subjects lost).

Figure 8 also shows that some trees that are objectively more complex than others seem empirically less complex. For example, according to Figure 8, tree 2.4 is empirically less complex than tree 2.3, and tree 2.3 is empirically less complex than tree 2.2. However, a simple right-tail t-test for difference in means (with different variance) shows that the percentage of subjects that failed in tree 2.4 is not significantly different than the percentage of subjects that failed either tree 2.3 or tree 2.2.

To represent all the possible pairwise comparisons of percentages of subjects who failed, we construct a heat map in Figure 9. Gray means that there is no statistically significant difference between these percentages, yellow means significant at the 10% level, orange means significant at the 5% level, and red means significant at the 1% level or less.

We compare each tree with all the trees that are empirically more complex. For example, take the first column in the heat map in Figure 9, which shows whether the empirical complexity of the least complex tree (2.4) is significantly different from the empirical complexity of other trees. This column shows that the difference in empirical complexity of tree 2.4 (a) is not significantly different from the empirical complexity of trees 2.3 and 2.2, (b) is significantly smaller at a 5% level than the empirical complexity of tree 3.2, and (c) is significantly smaller at the 1% level than the empirical complexity of any other tree.

Similarly, the heat map shows that tree 2.2.2.2.2 is significantly more complex than any other tree at the 1% level, except for tree 4.2.2.2, for which the difference in empirical complexity is significant at the 10% level.\textsuperscript{12}

While our definition of objective complexity allows us to compare 136 pairs of trees, the heat map presented in Figure 9 together with our definition of empirical complexity allow us to compare all 702 possible cases. The heat map reveals that some of the trees that are objectively less complex than others appear to be empirically more complex. This is the case, for example, of tree 2.3.2.2.

\textsuperscript{12}Since we always compare an empirically less complex tree with respect to a more complex one, the differences are always negative and as such we use the one-side p-value for the t-stats level of significance to construct the heat map.
which is empirically more complex than tree 2.4.2.2. We do not have as of now a rationalization for this result; however, comparing the heat map with respect to Figure 7, which shows the map of objective complexity, we find that a tree that is objectively more complex is empirically less complex at the 1% level of significance in only 8 out of 136 possible cases. At the same time, a tree that is objectively more complex is empirically more complex at the 1% level of significance in 97 out of 136 possible cases.

4.3 Empirical complexity: Driving forces

Our next objective is to analyze the main drivers of empirical complexity. This is important because empirical complexity can shed new light on most situations in which objective complexity is silent. For example, looking at the map of objective complexity (Figure 7), we observe that tree 2.2.2.2.2.2 is objectively more complex than only four of the alternative 26 trees we analyze. However, the heat map in Figure 9 shows that tree 2.2.2.2.2.2 is empirically more complex than all the other 26 trees we compare it to, and the results are significant at the 1% level or less in 25 out of the 26 cases.

As discussed before, there are two ways to increase the complexity of a tree. We can expand the number of rounds or we can expand the number of actions per round. To analyze the impact of increasing the number of rounds, we first group the trees by their number of rounds and calculate a weighted average (by the number of subjects) of the percentage of subjects who lose in trees within each group. Table 1 presents the groups of trees by number of rounds while the empirical complexity results are in Figure 10.

[Figure 10 about here.]

Figure 10 clearly shows that increasing the number of rounds increases the empirical complexity of a tree. Additionally, after the second round, adding an additional round increases the empirical complexity of a problem substantially. Moving from 3 rounds to 4 rounds more than triples our measure of empirical complexity. Adding two more rounds after the fourth round doubles the empirical complexity of the trees.

We now study the impact of increasing the number of actions per round. Since we already know
that the number of rounds increases empirical complexity, we present the results in two tables that allow us to separate the impact of the length and width dimensions. In Table 3, we present the results for a selected group of trees in which only one dimension changes at a time, either the length increases by one round or the width increases by one action on the first round.

Table 3 consists of two panels. Each nonempty cell in Panel (a) contains a tree while the associated cell in Panel (b) depicts the empirical complexity of that tree. We read Table 3 in the following way. Take a nonempty cell in Panel (a). A tree in a cell to the right is objectively more complex since a round with two actions has been added; this tree is longer. A tree in a cell below is also objectively more complex since one action has been added to the first round; this tree is wider.

Table 3 shows that the impact on the empirical complexity of adding a round surpasses the impact of adding an action. For example, for trees with the format $N1.N2$, adding an action increases the empirical complexity from 3.1% to 3.8%. However, adding a round ($N3 = 2$) increases the empirical complexity by more than double.

Similarly, for trees with the form $N1.N2.N3$, the impact of increasing the number of actions is much smaller than the impact of increasing the number of rounds. This pattern follows every cell in our table except for that corresponding to the tree 2.2.2.2. In this case, adding a round or adding an action reduced empirical complexity. This tree is one of the anomalies we observe in the database that we discussed previously.

Overall, it is clear that increasing the number of rounds increases empirical complexity by a large margin when compared to the impact on complexity of increasing the number of actions.

Now, we extend the previous analysis to use all trees in our experiment. For this purpose, in Table 4, for each set of trees with the same number of rounds, we divide them into a group with low (Min), medium (Med), and high (Max) number of actions per round. The difference between Tables 4 and 3 is that the former allows for the number of rounds to increase at any stage in the tree, and sometimes the number of actions might increase by more than one.

Within each group, the results are weighted by the number of subjects that played each tree. As with Table 3, the descriptions of the trees in each cell are presented in Panel (a) while the

[Table 3 about here.]
percentage of subjects who do not backward induct is depicted in Panel (b).

[Table 4 about here.]

The first row and the last two columns in Table 4 repeat the results from Table 3. Therefore, we focus on the first three columns and the last two rows (shaded in gray) in Table 4. The results corroborate what we found in Table 3: while increasing the number of actions might increase the empirical complexity of a tree marginally, increasing the numbers of rounds by one more than doubles or even triples the observed empirical complexity of the trees. In short, tree length, rather than tree width, is what drives empirical complexity.

5 Conclusions

Backward induction is the fundamental solution concept for finite dynamic games. However, as the vast literature shows, people do not seem to behave in accordance with backward induction. In this paper, we offer a novel explanation for the observed violations of backward induction, namely, tree structure.

We conduct a mobile experiment with 27 trees. We find that the percentage of subjects who violate backward induction (i.e., empirical complexity) could be as little as 2% and as high as 49%. While typical reasons, like subject’s limited cognitive skills, explain why these empirical complexity rates are not zero, it is necessary to rely on alternative explanations to understand why the empirical complexity differs across the trees.

The variation of complexity rate is mostly driven by the tree length (number of rounds). While adding actions at each round (and keeping the number of rounds constant) increases the percentage of subjects who violated backward induction, it is the addition of rounds (even with just two actions at each node) that makes the tree significantly more complex.

References


Figure 1: Tree A

```
    Ann
     /   \
    L     R
   /     /
AI α₁ β₁ AI α₂ β₂
  /     /
1,0 1,0 0,1 1,0
```
Figure 2: Tree B

![Tree B Diagram]

- Root: Ann
- Left Branch: L
  - Child 1: AI
    - Sub-Branch 1: \(\alpha_1\), \(\beta_1\), \(\gamma_1\)
    - Values: 1, 0, 1
  - Child 2: AI
    - Sub-Branch 2: \(\alpha_2\), \(\beta_2\), \(\gamma_2\)
    - Values: 1, 0, 1
- Right Branch: R
  - Child 1: AI
    - Sub-Branch 3: \(\alpha_3\), \(\beta_3\), \(\gamma_3\)
    - Values: 0, 1, 1
  - Child 2: AI
    - Sub-Branch 4: \(\alpha_4\), \(\beta_4\), \(\gamma_4\)
    - Values: 1, 0, 1
Figure 3: Tree C
Figure 4: Example of a tree in *Blues and Reds*. 

Figure 5: Tree 1
Figure 6: Tree 2
Figure 7: Pairwise comparison of trees in terms of objective complexity

This table presents the pairwise comparison of trees in accordance with objective complexity. Take a tree X from the vertical axis and a tree Y from the horizontal axis. If the symbol at the intersection of the Xth row and Yth column is \( \triangleright \), then \( X \trianglerighteq Y \) (X is objectively more complex than Y), and if the symbol is a dot, then it is not possible to objectively compare X and Y.
Figure 8: Ranking of trees by empirical complexity

This figure presents the ranking of trees by their empirical complexity measured as the percentage of subjects who do not behave in accordance with backward induction.
Figure 9: Heat map of empirical complexity

This table presents the pairwise comparison of trees in accordance with empirical complexity. Trees are ordered from the most complex (2.2.2.2.2.2) to the least complex (2.4). Colors indicate statistical significance of difference in percentages of subjects who failed (i.e., did not behave in accordance with backward induction). Red – significant at 1%, orange – significant at 5%, yellow – significant at 10%, and grey – not statistically significant.
Figure 10: Empirical complexity by the number of rounds

This figure presents the weighted average of the percentage of subjects who do not behave in accordance with backward induction in trees with 2, 3, 4, 5, and 6 rounds.
Table 1: Trees used in the experiment.

<table>
<thead>
<tr>
<th>2 rounds</th>
<th>3 rounds</th>
<th>4 rounds</th>
<th>5 rounds</th>
<th>6 rounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2</td>
<td>2.2.2</td>
<td>2.2.2.2</td>
<td>2.2.2.2</td>
<td>2.2.2.2</td>
</tr>
<tr>
<td>2.3</td>
<td>2.2.3</td>
<td>3.2.2.2</td>
<td>3.2.2.2</td>
<td>3.2.2.2</td>
</tr>
<tr>
<td>2.4</td>
<td>2.3.2</td>
<td>4.2.2.2</td>
<td>4.2.2.2</td>
<td>4.2.2.2</td>
</tr>
<tr>
<td>3.2</td>
<td>2.3.3</td>
<td>2.3.2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>3.2.2</td>
<td>2.4.2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.2.3</td>
<td>2.2.3.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.3.2</td>
<td>2.2.4.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.3.3</td>
<td>2.2.2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.2.2</td>
<td>2.2.2.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Number of subjects per tree.

<table>
<thead>
<tr>
<th>tree</th>
<th>number of subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.2</td>
<td>657</td>
</tr>
<tr>
<td>2.2.2.2</td>
<td>630</td>
</tr>
<tr>
<td>2.2.2.2.2</td>
<td>600</td>
</tr>
<tr>
<td>3.3</td>
<td>634</td>
</tr>
<tr>
<td>3.3.2</td>
<td>620</td>
</tr>
<tr>
<td>2.2.2.4</td>
<td>595</td>
</tr>
<tr>
<td>2.2</td>
<td>647</td>
</tr>
<tr>
<td>3.3.3</td>
<td>626</td>
</tr>
<tr>
<td>4.2.2.2</td>
<td>609</td>
</tr>
<tr>
<td>2.3.3</td>
<td>628</td>
</tr>
<tr>
<td>3.2.3</td>
<td>615</td>
</tr>
<tr>
<td>3.2.2.2</td>
<td>601</td>
</tr>
<tr>
<td>2.3</td>
<td>657</td>
</tr>
<tr>
<td>2.3.2</td>
<td>635</td>
</tr>
<tr>
<td>2.3.2.2</td>
<td>611</td>
</tr>
<tr>
<td>3.2</td>
<td>634</td>
</tr>
<tr>
<td>3.2.2</td>
<td>625</td>
</tr>
<tr>
<td>2.2.2.3</td>
<td>607</td>
</tr>
<tr>
<td>2.2.3</td>
<td>675</td>
</tr>
<tr>
<td>2.2.3.2</td>
<td>660</td>
</tr>
<tr>
<td>2.2.2.2.2</td>
<td>614</td>
</tr>
<tr>
<td>2.4</td>
<td>650</td>
</tr>
<tr>
<td>2.4.2.2</td>
<td>627</td>
</tr>
<tr>
<td>3.2.2.2.2</td>
<td>594</td>
</tr>
<tr>
<td>4.2.2</td>
<td>673</td>
</tr>
<tr>
<td>2.2.4.2</td>
<td>654</td>
</tr>
<tr>
<td>4.2.2.2.2</td>
<td>597</td>
</tr>
</tbody>
</table>
Table 3: Empirical complexity 1: Length and width.

These tables present the analysis of how tree length and width influence empirical complexity. For each cell in Panel (a) with a tree, there is an associated cell in Panel (b) with this tree’s empirical complexity.

Panel (a)

<table>
<thead>
<tr>
<th></th>
<th>2 rounds</th>
<th>3 rounds</th>
<th>4 rounds</th>
<th>5 rounds</th>
<th>6 rounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 actions</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>3 actions</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td>4 actions</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Panel (b)

<table>
<thead>
<tr>
<th></th>
<th>2 rounds</th>
<th>3 rounds</th>
<th>4 rounds</th>
<th>5 rounds</th>
<th>6 rounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 actions</td>
<td>3.1%</td>
<td>6.8%</td>
<td>33.3%</td>
<td>26.5%</td>
<td>49.3%</td>
</tr>
<tr>
<td>3 actions</td>
<td>3.8%</td>
<td>8.2%</td>
<td>25.1%</td>
<td>29.0%</td>
<td></td>
</tr>
<tr>
<td>4 actions</td>
<td>9.1%</td>
<td>26.8%</td>
<td>45.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Empirical complexity 2: Length and width.

These tables present the analysis of how tree length and width influence empirical complexity. For each cell in Panel (a) with a tree, there is an associated cell in Panel (b) with this tree’s empirical complexity.

Panel (a)

<table>
<thead>
<tr>
<th></th>
<th>2 rounds</th>
<th>3 rounds</th>
<th>4 rounds</th>
<th>5 rounds</th>
<th>6 rounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>Med</td>
<td>2.3</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td>Max</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Panel (b)

<table>
<thead>
<tr>
<th></th>
<th>2 rounds</th>
<th>3 rounds</th>
<th>4 rounds</th>
<th>5 rounds</th>
<th>6 rounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>3.1%</td>
<td>6.8%</td>
<td>33.3%</td>
<td>26.5%</td>
<td>49.3%</td>
</tr>
<tr>
<td>Med</td>
<td>3.0%</td>
<td>7.5%</td>
<td>26.7%</td>
<td>29.0%</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>4.2%</td>
<td>8.5%</td>
<td>21.7%</td>
<td></td>
<td>45.1%</td>
</tr>
</tbody>
</table>