UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Shades of Zero: Distinguishing impossibility from inconceivability

Permalink

https://escholarship.org/uc/item/5623p0kg

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Hu, Jennifer
Sosa, Felix Anthony
Ullman, Tomer D.

Publication Date

2024

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Shades of Zero: Distinguishing impossibility from inconceivability

Jennifer Hu[1]
jenniferhu@fas.harvard.edu
Kempner Institute
Harvard University

Felix Sosa[1]
fsosa@fas.harvard.edu
Department of Psychology
Harvard University

Tomer Ullman
tullman@fas.harvard.edu
Department of Psychology
Harvard University

Abstract

Eating onion ice cream is improbable, and levitating ice cream is impossible. But scooping ice cream using sadness is not just impossible: it is inconceivable. While prior work has examined the distinction between improbable and impossible events, there has been little empirical research on inconceivability. Here, we report a behavioral and computational study of inconceivability in three parts. First, we find that humans reliably categorize events as inconceivable, separate from probable, improbable, and impossible. Second, we find that we can decode the modal category of a sentence using language-model-derived estimates of subjective event probabilities. Third, we reproduce a recent finding that improbable events yield slowest response times in a possibility judgment task, and show that inconceivable events are faster to judge than impossible and improbable events. Overall, our results suggest that people distinguish the impossible from the inconceivable, and such distinctions may be based on graded rather than discrete judgments.

Keywords: impossibility; inconceivability; modal reasoning; language models; event knowledge; type errors

Introduction

Some things are impossible, but some things are more impossible than others. Levitating a feather with one's mind is impossible, but still easier or more probable than levitating a rock (Shtulman & Morgan, 2017; McCoy & Ullman, 2019). Such graded judgements of impossibility are the topic of ongoing study in cognitive science and cognitive development, with the general motivation that studying what makes things easier or harder in the imagination reveals people's understanding of everyday reality. Within this research direction, there have been different (though not mutually exclusive) explanations for what makes some things seem more impossible than others. These include causal violations (Shtulman & Morgan, 2017), violations of core knowledge and intuitive physics (McCoy & Ullman, 2019; Lewry, Curtis, Vasilyeva, Xu, & Griffiths, 2021), and moves across ontological hierarchies (Griffiths, 2015), among others.

But, just as there is a dividing line between the merely improbable and the impossible, there may be a category of events even more impossible than impossible. Levitating a feather with your mind is impossible in our world, but can still be imagined as occurring in a fictional world, and fit into our intuitive theories of possible worlds. By contrast, 'levitating a feather using the number five', or 'finding the square-root of dogs' are events that can't be evaluated or construed in any possible world. Borrowing from philosophy, we refer to such events as *inconceivable* (Gendler & Hawthorne, 2002).

One view is that in day-to-day decision making and reasoning, we judge whether things are possible or impossible based on whether we can conceive of a scenario in which they occur (Gendler & Hawthorne, 2002). For example, it is easy to conceive of a "red square", but not "a square that is not a square", and thus we might judge the former to be possible, while the latter is impossible. Of course, some things that we can conceive of might not be possible – we can imagine a world where levitation exists, while knowing that it is impossible given the physical laws of the natural world. While the relationship between conceivability and possibility has been the topic of philosophical research, there has been little empirical and computational cognitive science study of inconceivability. A related line of work in cognitive development has investigated people's distinction between the *impossible* and the merely *improbable* (Shtulman & Carey, 2007; Shtulman, 2009; Goulding, Khan, Fukuda, Lane, & Ronfard, 2023), building upon the assumption that there is a dividing line between physical impossibilities (e.g., walking on water) and "pseudo-impossibilities" (e.g., growing a beard down to one's toes). However, it remains poorly understood whether inconceivability and impossibility also form meaningfully distinct modal categories, or whether people treat inconceivability as simply an instance of impossibility.

Here we report a behavioral and computational study of inconceivability, in three parts. In the first part, we examined whether people readily and reliably distinguish inconceivable events from other modal categories (probable, improbable, and impossible). Such a categorization test follows the logic of cognitive development studies used to conclude young children do not distinguish improbable from impossible events (Shtulman & Carey, 2007). We found that people are highly consistent in their categorizations, suggesting these categories are easily distinguished from each other.

People's ready distinction between the impossible and inconceivable raises the question of how such a distinction is made: is it a difference in *kind* or a difference in *degree*? In other words, are inconceivable events processed in a qualitatively different manner than impossible events, or are they just really, really impossible? One hypothesis in support of the first view is that inconceivability is the result of the mind hitting a category error (Magidor, 2009; Ryle, 1949),

[1]Equal contribution.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Prefix	Probable	Improbable	Impossible (Physics)	Impossible (Magic)	Inconceivable (Semantics)	Inconceivable (Syntax)
Baking a cake using	an oven	an airfryer	a refrigerator	superpowers	anger	grasp
Chilling a drink using	ice	snow	fire	pixies	tomorrow	at
Washing your hair with	shampoo	detergent	air	mermaids	minutes	though
Drawing a picture using	a pencil	lipstick	a mountain	magic	a smell	always

Table 1: Sample stimuli used in our experiments.

a point at which people's mental model breaks down. Such a process would parallel the "type errors" encountered by type-based computer programs. For example, the expression 'square_root(45)' will be evaluated by most computer programs, but trying to evaluate 'square_root('dog')' will be rejected as a type violation, as 'dog' is simply not the kind of thing you can apply the square-root program to. Some researchers have proposed that types, which enforce the expected inputs and outputs of a program, may form the basis of mental computation across domains and behaviors (Morales, 2018; Sosa & Ullman, 2022). The argument would then go that inconceivable events are categorized on the basis of type errors, while impossible events are not. However, a different hypothesis to the type-error process is that people have a single, graded notion of probability where all modal events exist on a spectrum, including the impossible and inconceivable. Modal categories could then be read out by defining a straightforward transformation on top of the underlying probabilities – for example, by defining thresholds on probability values (improbable events are 1-in-100, impossible are 1-in-a-million, inconceivable is 1-in-a-trillion). This hypothesis may accord with proposals that selectional restrictions are based on statistical associations (Resnik, 1993, 1996).

In the second part, we contrast the hypotheses that people's distinction between impossible and inconceivable are one of kind or degree. We examined whether events that vary in their modality can be distinguished through their subjective probability, as estimated by string probabilities from language models, which can be treated as purely probabilistic models of language generation. If people categorize inconceivable and impossible events as separate, while both are assigned the same near-zero probability, it would suggest the distinction is not made on the basis of probability. However, we find that we can decode the modal category of a sentence using a model's log-probabilities, suggesting these modal categories can be distinguished using just probability.

In the third and final part, we examined the response time (RT) associated with judging the possibility of an event, as a measure of processing difficulty (cf. Goulding et al., 2023), and a potential behavioral signature that distinguishes the impossible and inconceivable. Using our novel materials, we reproduce Goulding et al.'s finding that improbable events take the longest time to judge. In addition, we found that inconceivable events are substantially *faster* to judge than impossible or improbable events.

Overall, our findings show that people do distinguish the impossible from the inconceivable, both by a direct categorization and by an indirect processing difficulty measure. We find that this distinction (and modal distinctions in general) can be decoded from probability as captured by language models, supporting the view that such judgments in people may be based on graded rather than discrete judgments. In the discussion, we consider limitations and additional steps needed to tease the graded and discrete options apart.

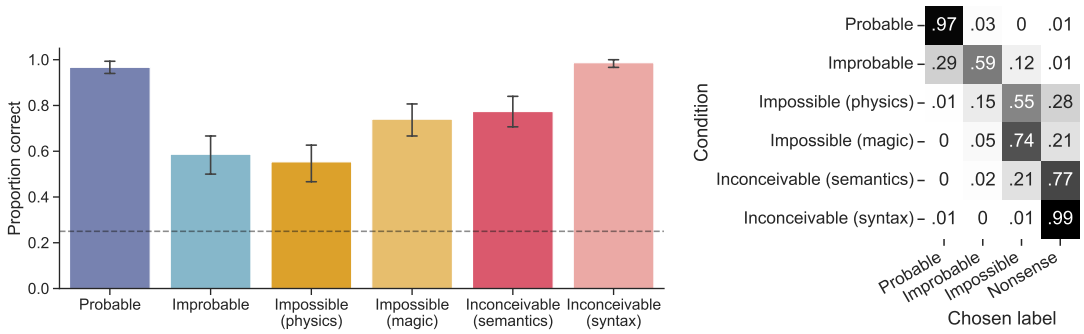
Experiment 1: Can people reliably categorize inconceivable events?

In our first experiment, we ask whether humans consistently categorize events as probable, improbable, impossible, or inconceivable. Our main questions are whether humans categorize events in a way that is consistent with the underlying coding of conditions, and what kinds of errors are made.

Stimuli We manually constructed a set of 30 items designed to cover several modal categories (see examples in Table 1). Each item consists of a shared prefix denoting a commonplace event with a transitive verb and object, as well as the beginning of a phrase describing how the event occurs (e.g., "Baking a cake using"). No explicit subject is specified. Each item prefix is associated with six candidate continuations, each of which completes the phrase describing how the event occurs. These continuations reflect different types of modal relationships to the prefix. Since the region that modulates the modal category of the event occurs at the end of the phrase, these stimuli are well-suited to test autoregressive (i.e., "left-to-right") language models, which condition on preceding context to predict the subsequent tokens.

The **Probable** and **Improbable** conditions reflect events that are possible given the physical laws of the real world. Probable continuations are prototypical or highly expected (e.g., "baking a cake using an *oven*", or "washing your hair with *shampoo*"). Improbable continuations are unconventional, but do not violate any physical or social constraints based on norms in the United States (e.g., "baking a cake using an *airfryer*", or "washing your hair with *detergent*").

We also considered two **Impossible** conditions: one where events are impossible because of physical constraints (ImpossiblePhysics), and one where events are impossible because of magic (ImpossibleMagic). For example, while baking a cake using a *refrigerator* or using *superpowers* are both physically impossible in the real world, the former is impossible



(a) (b)

Figure 1: (a) Proportion of trials in Experiment 1 where human responses matched the underlying condition coding. Dashed line indicates chance (25%). Error bars indicate bootstrapped 95% CI. (b) Response rates in each condition of Experiment 1. Cell annotations are rounded to 2 decimal places for visualization purposes, so row values may appear to not sum to 1.

due to thermodynamics whereas the latter appeals to a magical concept that does not exist. We investigate magic-based impossibility because we people might reason about these scenarios based on intuitive physics (e.g., levitating rocks versus levitating feathers; Shtulman & Morgan, 2017).

Finally, we considered two types of **Inconceivable** conditions: one where events constitute category errors (InconceivableSemantic), and one where the stimulus has a syntactic violation (InconceivableSyntactic). The InconceivableSemantic condition completes the prefix with an abstract noun that renders the event inconceivable (e.g., "chilling a drink using *tomorrow*"). The InconceivableSyntactic condition features a continuation that makes the full phrase ungrammatical, typically by featuring a verb, adverb, or preposition instead of the expected noun (e.g., "chilling a drink using *at*"). This condition serves as a baseline, since we expect both humans and language models (LMs) to recognize these expressions as ill-formed. In the current experiment, this means that we expect humans to reliably distinguish these items from the Probable, Improbable, and Impossible items, all of which are grammatically well-formed. In Experiment 2, we expect LMs to reliably assign lower probabilities to these continuations, as shown in prior LM testing (Marvin & Linzen, 2018; Hu, Gauthier, Qian, Wilcox, & Levy, 2020; Warstadt et al., 2020).

Methods We recruited $N = 30$ participants on Prolific, with a self-reported native language of English. Participants were compensated at an hourly rate of $12. Each participant saw each of the 30 items, one item per trial. On each trial, participants saw a prefix and a continuation in one of the six conditions (e.g., "Baking a cake using an oven"). Their task was to categorize the stimulus into one of four categories (by pressing keyboard buttons): "probable", "improbable", "impossible", or "nonsense".[1] After responding on each trial, participants saw a screen saying "Your response has been logged" for 1 second. Each participant saw an equal number of trials in each condition (i.e., 5 trials for each of the 6 conditions). The order of trials and conditions was randomized. Prior to the experiment, participants were familiarized with definitions and examples of each category, and were required to correctly complete 8 practice trials.

Results Figure 1a shows accuracy for each condition, or the proportion of trials where human responses matched the underlying condition coding. Participants achieved near-ceiling accuracy in the Probable and InconceivableSyntax conditions, lower accuracy in the ImpossibleMagic and InconceivableSemantics conditions, and lowest (but above-chance) accuracy in the Improbable and ImpossiblePhysics conditions.

Errors in these conditions were not random, but instead reflected structured patterns (Figure 1b). For example, people were most likely to mislabel the Improbable items as "probable" (29%). This seems perfectly reasonable, as the distinction between Probable and Improbable is a matter of degree: both types of events are possible, but differ in the magnitudes of their probabilities, which can vary based on each individual's experiences and environments. For example, a culinary student who specializes in creative airfryer recipes might be more likely to label "baking a cake using an airfryer" as "probable" than the average home baker. Interestingly, there were extremely few cases of mislabeling Probable items as "improbable" (3%). We speculate that this asymmetry could be attributed to the quality of the Probable stimuli: they reference common, prototypical events that are likely shared across the life experience of our participants.

We also observe a slight tendency to mislabel Improbable items as "impossible", and vice versa. Improbable items were labeled as "impossible" 12% of the time, and ImpossiblePhysics items were labeled as "improbable" 15% of the time. One potential explanation is that modal judgments of (im)possibility are subject to personal experience in the same way judgments of (im)probability are. Notably, the depen-

[1]We used the label "nonsense" instead of "inconceivable" to avoid jargon, while capturing a similar intuition.

dence of (im)possibility on life experience has been observed in children, where younger children are more likely they are to judge a highly improbable event as impossible (Shtulman & Carey, 2007). While the majority of responses correctly label Impossible* items as "impossible" and Inconceivable* items as "nonsense", we also observe some confusability between impossibility and inconceivability. 28% of responses mislabeled ImpossiblePhysics as "nonsense", and 21% mislabeled InconceivableSemantics as "impossible". These patterns suggest that, while people readily distinguish inconceivable from impossible (and other modal categories), the boundaries between these categories may be graded and depend on individuals' subjective experiences.

Experiment 2: Do probabilities distinguish conceivable and inconceivable events?

Next, we ask whether events with different modalities (probable, improbable, impossible, or inconceivable) can be distinguished based on a single graded quantity: their subjective probability of occurring. In order to estimate these probabilities, we leverage state-of-the-art neural network language models (LMs). LMs are trained with the objective of predicting sequences of tokens, by which they may implicitly learn the latent properties of the world that make certain linguistic expressions more or less likely. Since people communicate about events (McRae & Matsuki, 2009) and observations (Louwerse, 2011, 2018), it may be reasonable to expect that language itself contains structured information about the world. Indeed, prior work has shown that LMs capture important aspects of commonsense and world knowledge (Chang & Bergen, 2023), such as the distinction between possible and impossible events (Kauf et al., 2023), and the structure of perceptual spaces (Abdou et al., 2021; Patel & Pavlick, 2022). The autoregressive next-token-prediction objective used to train LMs also has connections to real-time language comprehension in humans: psycholinguistic studies have demonstrated that humans engage in prediction about upcoming linguistic content (Altmann & Kamide, 1999; Levy, 2008; Smith & Levy, 2013) and use event knowledge to update their expectations (McRae & Matsuki, 2009; Bicknell, Elman, Hare, McRae, & Kutas, 2010; Matsuki et al., 2011). It is therefore plausible that LMs may learn structured information about the world (and events occurring in the world) in service of optimizing the objective of next-word prediction. Building upon these arguments, we use LMs as a tool for computing fine-grained estimations of event probabilities, with the assumption that string descriptions of more probable events will be assigned higher probability.[2]

Methods We use the same stimuli as in Experiment 1 (Table 1). To estimate how (un)expected an event is, we measure the surprisal S (Hale, 2001; Levy, 2008), or negative log probability, of the continuation c conditioned on its prefix p:

$$S(c,p) = -\log P(c|p) \quad (1)$$

If a model has learned to represent event probabilities in a way that conforms to the normative coding of our stimuli, then it should assign lowest surprisal to Probable continuations, higher surprisal to Improbable continuations, and highest surprisal to Impossible and Inconceivable continuations.

By comparing the surprisal of different continuations conditioned on the same prefix, there is the potential confound of observing differences driven by frequency effects. For example, we might observe $S(\text{an airfryer}|\text{Baking a cake using}) > S(\text{an oven}|\text{Baking a cake using})$, simply because "an aifryer" (the Improbable continuation) occurs less frequently than "an oven" (the Probable continuation) in written text. One way to address this would be to hold the continuation constant while varying the prefix across comparisons, as is done in other targeted evaluation of LMs. However, this is infeasible for our materials: for example, the continuations in the ImpossibleMagic condition (like "superpowers") could never be physically possible. Therefore, we instead validate our materials by comparing the n-gram frequency counts of each continuation across conditions.

We evaluated two open-source large language models: Llama-2-7B (Touvron et al., 2023) and Mistral-7B-v0.1 (Mistral-7B; Jiang et al., 2023). Both models are autoregressive Transformers with 7 billion parameters. Llama-2-7B was trained on 2 trillion tokens of internet text, whereas the training data details of Mistral-7B are unknown.

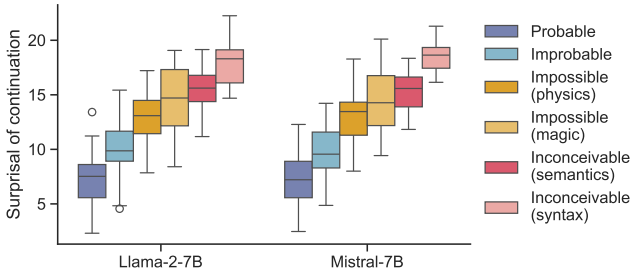


Figure 2: Surprisal (negative log probability) values assigned by language models to continuations in each condition.

Results Figure 2 shows the surprisal values for continuations in each condition. For both of our tested models, the condition-level surprisal averages (means and medians) are ordered in the following way (mirroring the ordering of the x-axis): Probable < Improbable < ImpossiblePhysics < ImpossibleMagic < InconceivableSemantics < InconceivableSyntax. A one-sided Kolmogorov-Smirnov test for these pairwise comparisons revealed significant differences between the distributions, except for ImpossibleMagic

[2]This assumption has limitations: for example, due to reporting biases, it is possible that text corpora may overestimate the occurrence of unlikely or impossible events (Gordon & Van Durme, 2013). We return to this issue in the Discussion.

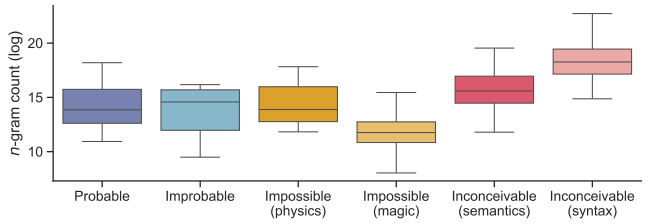


Figure 3: n-gram log frequency counts (estimated from Google Books) of continuations in each condition.

$<$ InconceivableSemantics.[3] This suggests that the modal categories from our materials can be distinguished in string probability space (with the potential exception of ImpossibleMagic and InconceivableSemantics). This contrasts with the findings of Kauf et al. (2023), who find that models struggle to differentiate between likely and unlikely events.

As a control, we wanted to ensure that the differences in surprisal values across conditions were not explained by differences in frequencies. We estimated the (unconditional) n-gram frequency of the continuations using the Google Books corpus. Figure 3 shows log frequency counts across conditions. The frequencies do not significantly differ across the Probable, Improbable, and ImpossiblePhysics conditions, and are in fact higher in the Inconceivable* conditions (which, if anything, would bias surprisal values to be lower in the Inconceivable* conditions relative to the others). This suggests that it is unlikely that n-gram frequency statistics are driving the surprisal orderings reported in Figure 2.

Experiment 3: How do people judge the possibility of inconceivable events?

In our third experiment, we measured reaction times (RTs) in a task where people are asked to judge whether an event is possible or not possible. This is a simpler task than the categorization tested in Experiment 1, and matches the tasks used in prior studies of RT across modal categories. Goulding et al. (2023) find that RTs are highest for improbable events, and lower for ordinary (here, what we call "probable") and impossible events. This experiment serves as a replication of this finding using a new set of materials, and also contributes new data about RT patterns for judging inconceivable events.

Stimuli We used the same set of stimuli as in Experiment 1, except we only kept items where at least 75% of ratings in Experiment 1 agreed with the underlying condition coding (averaged across conditions). This resulted in 18 items for Experiment 3. We also made some minor changes to improve the clarity of the items.

Methods In order to get a robust estimate of RTs, we recruited a large sample ($N = 299$) of participants on Prolific, with a self-reported native language of English. Participants were compensated at an hourly rate of $12. Each participant saw each of the 18 items, one item per trial. On each trial, participants saw a question of the form "Could someone [PREFIX] [CONTINUATION]?", where the continuation comes from one of the six conditions (e.g., "Could someone bake a cake using an oven?"). Their task was to respond "Yes" or "No" by pressing keyboard buttons '1' or '0'. After responding on each trial, participants saw a screen saying "Your response was logged" for 1 second. Each participant saw an equal number of trials in each condition (i.e., 3 trials for each of the 6 conditions). Prior to the experiment, participants were required to correctly complete 8 practice trials.

Response times were measured in milliseconds using browser-based jsPsych, which has been empirically validated (Reimers & Stewart, 2015) and produces RT measurements similar to those by standard psychophysics software (de Leeuw & Motz, 2016). We preprocessed the data in the following way. First, for basic quality control, we only kept participants who achieved an accuracy of at least 80% and passed the comprehension check within 3 attempts. Then, we only kept participants with reasonable variance in RT (to avoid people who always respond as quickly as possible) by removing participants whose standard deviation in RT was at least two standard deviations away from the mean standard deviation. After performing these exclusions, we had data from 247 participants. Finally, we normalized RTs within each participant, and removed RTs that were at least three standard deviations away from the participant mean.

Results Figure 4 shows normalized response time means across the six tested conditions. First, we reproduce the finding from Goulding et al. (2023) using our materials: RTs are the highest in the Improbable condition, lowest in Probable, and second highest in ImpossiblePhysics. This pattern is consistent with the model proposed by Shtulman (2009), where people first generate a modal intuition, and then reflect on this intuition (e.g., by simulating the event; Shtulman & Carey, 2007) before finally reaching a modal judgment. On this view, people develop modal intuitions by searching through memory to retrieve knowledge or experience that is similar to the event in question. This process takes less time for events that are frequently encountered (i.e., in the Probable condition) or clearly beyond the realm of normal experience (i.e., in the Impossible condition), but takes more time for events that are unfamiliar but do not immediately violate any physical or causal principles (i.e., in the Improbable condition).

In addition, we find that RTs are substantially *faster* in the InconceivableSemantics condition than in the Improbable and Impossible conditions. One explanation for why inconceivable events are judged so quickly, relative to improbable and impossible events, is that it is easy to recognize them as inconsistent in some ontological way. For example, "locking a door with a day" might be so clearly nonsensical that people bypass any further reflection about the event.

[3] We additionally verified that the patterns observed in Figure 2 held at the item-level (i.e., across conditions for a given item prefix).

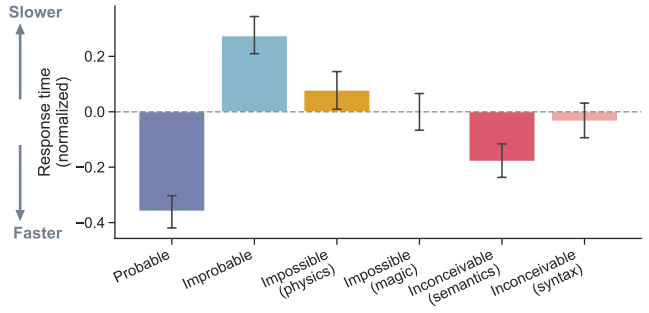


Figure 4: Mean normalized response time for each condition in Experiment 3 (binary judgment task).

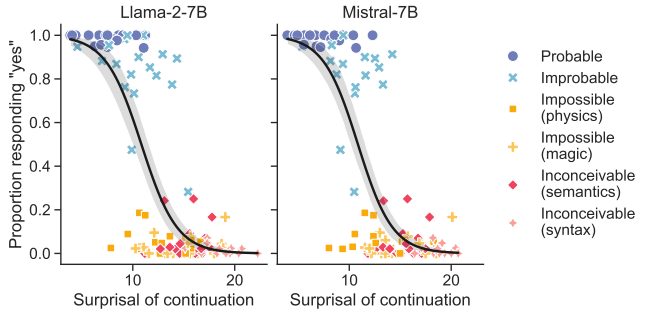


Figure 5: Proportion of "yes" responses (i.e., affirming the event to be possible) in Experiment 3 versus model-derived surprisal of continuation.

Experiment 3 also served as a way to validate whether model-derived surprisal values capture people's subjective judgments of event probabilities. If surprisals do approximate these subjective judgments, then we would expect that the surprisal of a continuation should predict whether people tend to affirm the possibility of the event. Figure 5 shows that the proportion of trials where people affirmed the event (i.e., judged it to be possible) does indeed decrease as the surprisal increases. We take this to suggest that, within the scope of our experiments, model-derived surprisals do a decent job of capturing people's subjective judgments of event probability.

Discussion

Beyond probable, improbable, and impossible, there is the inconceivable. While a great deal of recent research has cognitively studied the distinctions between and within improbable and impossible events, far less work has empirically or computationally examined people's reasoning about inconceivable events. Here, we investigated inconceivable events as a distinct mental category, and found that people reliably distinguished inconceivable events from other modal categories, including impossible ones, and showed different associated processing times as measured by RT for this category. These findings expand on an ongoing research program in cognitive science and cognitive development that examines people's varying judgements of impossible events, and their distinction from the merely improbable.

While people distinguish the impossible and the inconceivable, this distinction may be one of degree, or of kind. To examine these different options, we used language models to estimate event probabilities for events of different modalities. We found that modal event categories can in principle be distinguished by the probabilities associated with them. Such a finding lends support to the possibility that the impossible/inconceivable distinction (and other modal distinctions) may be based on a graded continuum. However, it remains an open possibility that people do in fact process the inconceivable in a different way than the impossible.

One specific proposal of a difference in kind between impossibility and inconceivability is that judgements of inconceivability follow the experience of a category error, which are analogous to type errors in typed computer programs (Sosa & Ullman, 2022). In much the same way that a computer program for calculating square-roots expects a number and would throw an error when encountering a string, the mind may expect 'baking a cake using a' to be followed by a *physical object*, and throw an error if it is followed by the wrong type (e.g. "the number three"). A promising direction for future work is to use inconceivability as a way to investigate potential behavioral signatures of type-based reasoning.

One limitation of our study is using LM-derived string probabilities to estimate the probability of an event. As discussed in Experiment 2, this methodology rests on an assumption that the likelihood of using a particular linguistic expression to describe an event is a proxy for how frequently the event occurs. This assumption faces at least two challenges. The first challenge, reporting bias, affects possible events. Low-likelihood events may be over-represented in language, since people are incentivized to talk about them, whereas high-likelihood events may be underrepresented in language, since there is little communicative benefit of talking about them (Sorower et al., 2011; Gordon & Van Durme, 2013). Reporting bias is most likely to skew models' probability estimates of Probable and Improbable events, as events with sufficiently low probabilities (e.g., Impossible and Inconceivable) might occur so infrequently that they are hardly described in language at all. In our study, we do find that LM probabilities reflect the expected ordering of Probable $>$ Improbable, suggesting that modern LMs can overcome reporting biases to some extent (cf. Shwartz & Choi, 2020). The second challenge affects non-possible events. There may be cases where a particular string completion makes an event impossible, but is still highly predictable given the context. For example, given "The wizard plucked a scale from his fire-breathing", the completion "dragon" might be extremely likely, even though dragons do not exist. This is not a major concern for the materials here, since the prefixes describe commonplace physical events, where completions that make the event impossible would likely only occur in highly specific contexts (and would thus be assigned low probability).

Acknowledgments

This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. The experiments were supported by a gift from the Harvard University Hodgson Memorial Fund.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021, November). Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 109–132). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.conll-1.9

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. doi: 10.1016/S0010-0277(99)00059-1

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*(4), 489–505. doi: 10.1016/j.jml.2010.08.004

Chang, T. A., & Bergen, B. K. (2023). Language Model Behavior: A Comprehensive Survey. *Computational Linguistics*, 1–55.

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12.

Gendler, T., & Hawthorne, J. (Eds.). (2002). *Conceivability and Possibility*. New York: Oxford University Press.

Gordon, J., & Van Durme, B. (2013). Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction* (pp. 25–30). New York, NY, USA: Association for Computing Machinery.

Goulding, B. W., Khan, F., Fukuda, K., Lane, J. D., & Ronfard, S. (2023). The development of modal intuitions: A test of two accounts. *Journal of Experimental Psychology: General*.

Griffiths, T. L. (2015). Revealing ontological commitments by magic. *Cognition*, *136*, 43–48.

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020, July). A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1725–1744). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.158 doi: 10.18653/v1/2020.acl-main.158

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., … Sayed, W. E. (2023). *Mistral 7B*. arXiv.

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., … Lenci, A. (2023). Event Knowledge in Large Language Models: The Gap Between the Impossible and the Unlikely. *Cognitive Science*, *47*(11), e13386.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126 – 1177.

Lewry, C., Curtis, K., Vasilyeva, N., Xu, F., & Griffiths, T. L. (2021). Intuitions about magic track the development of intuitive physics. *Cognition*, *214*.

Louwerse, M. M. (2011). Symbol Interdependency in Symbolic and Embodied Cognition. *Topics in Cognitive Science*, *3*(2), 273–302. doi: https://doi.org/10.1111/j.1756-8765.2010.01106.x

Louwerse, M. M. (2018). Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps. *Topics in Cognitive Science*, *10*(3), 573–589. doi: https://doi.org/10.1111/tops.12349

Magidor, O. (2009). Category mistakes are meaningful. *Linguistics and Philosophy*, *32*(6), 553–581.

Marvin, R., & Linzen, T. (2018, October). Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1192–1202). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1151 doi: 10.18653/v1/D18-1151

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011, July). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 913–934. doi: 10.1037/a0022964

McCoy, J., & Ullman, T. (2019). Judgments of effort for magical violations of intuitive physics. *PLOS ONE*, *14*(5).

McRae, K., & Matsuki, K. (2009, November). People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible. *Language and Linguistics Compass*, *3*(6), 1417–1429. doi: 10.1111/j.1749-818X.2009.00174.x

Morales, L. E. (2018). *On the representation and learning of concepts: Programs, types, and Bayes*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Patel, R., & Pavlick, E. (2022). Mapping Language Models to Grounded Conceptual Spaces. In *International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=gJcEM8sxHK

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*(2), 309–327.

Resnik, P. (1993). Semantic Classes and Syntactic Ambiguity. In *Human Language Technol-*

ogy: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993. Retrieved from https://aclanthology.org/H93-1054

Resnik, P. (1996). Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, *61*, 127–159.

Ryle, G. (1949). *The Concept of Mind*. University of Chicago Press.

Shtulman, A. (2009, July). The development of possibility judgment within and across domains. *Cognitive Development*, *24*, 293–309. doi: 10.1016/j.cogdev.2008.12.006

Shtulman, A., & Carey, S. (2007). Improbable or Impossible? How Children Reason About the Possibility of Extraordinary Events. *Child Development*, *78*(3), 1015–1032.

Shtulman, A., & Morgan, C. (2017). The explanatory structure of unexplainable events: Causal constraints on magical reasoning. *Psychonomic Bulletin & Review*, *24*(5), 1573–1585.

Shwartz, V., & Choi, Y. (2020). Do Neural Language Models Overcome Reporting Bias? In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6863–6870). International Committee on Computational Linguistics.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302 – 319.

Sorower, M., Doppa, J., Orr, W., Tadepalli, P., Dietterich, T., & Fern, X. (2011). Inverting Grice's Maxims to Learn Rules from Natural Language Extractions. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 24). Curran Associates, Inc.

Sosa, F. A., & Ullman, T. (2022). Type theory in human-like learning and inference. In *Beyond Bayes: Paths Towards Universal Reasoning Systems Workshop at the International Conference on Machine Learning.*

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models.* arXiv.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, *8*.