Making Sense of Nonsense

Jennifer Hu (jenniferhu@fas.harvard.edu)
Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Boston, MA 02134, USA

Felix Sosa (fsosa@fas.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA 02138, USA

Tomer Ullman (tullman@fas.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA 02138, USA
Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Boston, MA 02134, USA

Abstract

Some impossible things are more impossible than others. Magically levitating a feather seems easier than levitating a rock, even though both are impossible in the real world. But within the things that are inconceivable—e.g., "the number 13 writing a play" or "a girl being a prime number"—are some things *more inconceivable* than others? We first established that people have graded, systematic judgements of the likelihood of inconceivable and nonsense sentences (Experiment 1). We then examined two hypotheses as to how people make such judgments: the ease of a metaphorical interpretation (Experiment 2), and how difficult it is to transform a nonsense statement into a sensible one, as measured by distance in a type hierarchy (Experiment 3). We found that graded judgments of inconceivability are not captured by metaphorizability, but do correspond to a measure of distance in a type hierarchy. Our results suggest that inconceivability is graded, and the perceived likelihood of an inconceivable event may be a product of one's ontology of the world.

Keywords: impossibility; inconceivability; category errors; type errors; modal reasoning; metaphor

Introduction

Some things are impossible, and some don't make any sense at all. If a child tells you they made friends with a dragon, you can safely write it off as a product of their imagination, as you know that to be impossible. But if the child said they made friends with the number 5, that would be ludicrous in an altogether different way, which we refer to as "nonsense" or "inconceivable". Impossible things like dragons don't exist in our world, but they can still be imagined as existing in a possible world (Lewis, 1986). By contrast, inconceivable things like "making friends with the number 5" or "hanging a coat on a yawn" cannot be literally evaluated or construed in any possible world (Gendler & Hawthorne, 2002).[1] Recent empirical work backs up this philosophical distinction, showing that people systematically distinguish between the impossible and the inconceivable (Hu, Sosa, & Ullman, 2025).

For impossible things, research has shown that people think some things are more impossible than others. For example, using magic to levitate a feather seems somehow easier than levitating a boulder (Shtulman & Morgan, 2017; McCoy & Ullman, 2019), even though both are impossible. Research has also examined *why* people think this way, motivated by the notion that people generally use their understanding of the real world to construct and evaluate imaginary world (Byrne, 2007), and so examining people's evaluation of imaginary worlds can tell us how they think about the real world. There have been a variety of (overlapping) explanations for these graded judgments of impossibility, including moves across ontological hierarchies (Griffiths, 2015), causal violations (Shtulman & Morgan, 2017), and violations of core knowledge (McCoy & Ullman, 2019; Lewry, Curtis, Vasilyeva, Xu, & Griffiths, 2021).

While impossible things (like using magic to levitate a feather) can be imagined in a possible world, this is not true for inconceivable things, by definition. People may evaluate the likelihood of an impossible event by instantiating it in a possible world model, but inconceivability is a failure to get the world model off the ground at all. Such events or statements seem more like "category mistakes" or "type errors" (Magidor, 2009, 2017; Sosa & Ullman, 2022), in which the truth or likelihood of the output of some computation cannot be achieved, as the computation itself returns an error. Again by definition, category errors are not meant to be graded.

And yet, in principle it is possible that people's sense of gradedness extends from the impossible to the inconceivable. That is, beyond the fact that people both intuitively and formally distinguish *impossible* from *nonsense* as categories, they may think that some nonsensical things are even *more* nonsensical than others. If they do, then just like it is worthwhile to consider the cognitive representations that underlie the gradedness of the impossible, it is worthwhile considering those that underlie the gradedness of the inconceivable.

Here, we do two things. First, we examine empirically whether people have consistent graded judgments of the inconceivable (Experiment 1). After establishing that they do, we explore two different, non-exhaustive hypotheses as to what cognitive mechanisms this gradedness is based on (Experiment 2 and 3).

To spell this out further: in Experiment 1, we ask whether people judge some inconceivable events as more likely to occur than others. We used stimuli from a previous study of impossibility and inconceivability, and evaluated whether people ranked inconceivable event descriptions in a systematic way. We reproduced the prior finding of systematic agreement within *impossible* events, but above and beyond this we found evidence for systematic agreement regarding *in-*

[1] One may already object that transformations of an event *can* make sense of it, like supposing "the number five" refers to an animated character in the *shape* of the number 5, or that the child means they solved a math problem. We consider such "coercions" later.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

ID	Stimulus	Mean rank
a	lifting a box using foam	2.73
b	drying your hands using a liquid	5.13
c	crushing a soda can using a tissue	5.72
d	chopping a carrot using a rag	5.81
e	carving a statue using a sock	6.45
f	boiling water using a refrigerator	6.73
g	starting a fire using milk	7.00
h	smashing a pumpkin using a leaf	7.01
i	baking a cake inside a freezer	7.05
j	writing a letter using a snowflake	7.11
k	storing dishes inside a thimble	7.83
l	chilling a drink using fire	8.01
m	splitting a log using a feather	8.11
n	turning an apple into a cherry	8.99
o	taking a vacation to the sun	11.31

Table 1: Stimuli in Impossible condition of Exp 1, sorted by ranking, from most likely (top) to least likely (bottom).

ID	Stimulus	Mean rank
a	docking a boat using a lesson	3.73
b	breaking a coconut using a texture	4.71
c	smashing a pumpkin using a number	5.62
d	building a house out of language	5.98
e	popping a balloon using a story	6.59
f	hiding a gift inside a gasp	7.14
g	displaying books on a sigh	7.41
h	rinsing fruit using a giggle	7.41
i	melting a chocolate bar using a promise	7.56
j	painting a wall using a whisper	7.61
k	tightening a bolt using a dream	7.85
l	squeezing juice from a mumble	8.47
m	digging a hole using a yawn	8.48
n	packing a suit inside a groan	8.80
o	growing flowers inside a sneeze	8.86

Table 2: Stimuli in Inconceivable condition of Exp 1, sorted by ranking, from most likely (top) to least likely (bottom).

conceivable: some descriptions of inconceivable events are consistently ranked as "more likely" than others.

We then explore two potential hypotheses for what might drive graded rankings of inconceivable events. In Experiment 2, we test the hypothesis that items for which it is easier to find a figurative or metaphorical meaning are perceived as more likely. We do this by asking people how good of a metaphor each item is. While we do find systemic agreement across people regarding which inconceivable statements are "better" metaphors, these scores are not at all predictive of the rankings from Experiment 1. In Experiment 3, we go on to test the hypothesis that people's graded interpretation of inconceivable events is informed by an ontology or world model, similar to the type systems found in programming languages (c.f. Sosa & Ullman, 2022). Using a novel set of stimuli, we find that rankings are predicted by the similarity structure between types. That is, an inconceivable event involving a type violation with two similar types (e.g., animate entity vs. inanimate entity) is judged as more likely than an inconceivable event involving a type violation with two distant types (e.g,. animate entity vs. numeral).

Our results suggest that just as not all impossible things are equally impossible, not all inconceivable things are equally inconceivable. Furthermore, people might make sense of nonsense using graded expectations about the components of the nonsense under question, in a way that mirrors how modern type systems are used to define and interpret programs.

Experiment 1: Establishing graded inconceivability

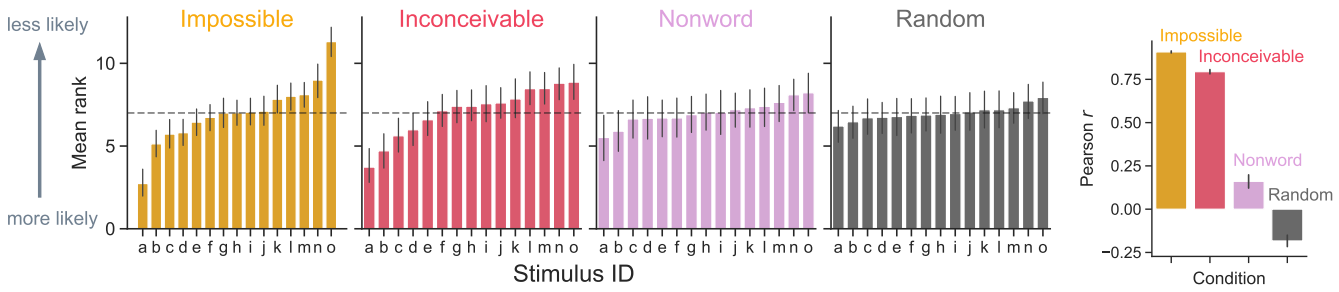
Stimuli We used a ranking measure similar to previous approaches for studying gradedness in impossibility. We used three sets of stimuli, each with 15 items: *Impossible*, *Inconceivable*, and *Nonword*. The three stimulus sets are shown in Table 1, Table 2, and Table 3, respectively.

The Impossible and Inconceivable event descriptions were taken from Hu et al. (2025). In both conditions, the event descriptions were short English phrases of the form "[verb]ing [noun phrase] [prepositional phrase]", where the prepositional phrase described the manner of the event. The Impossible event descriptions involved violations of basic physical principles in the real world: for example, "smashing a pumpkin using a leaf" or "chilling a drink using fire". In contrast, the Inconceivable event descriptions featured an abstract or non-physical noun when the event expected a concrete noun: for example, "displaying books on a sigh" or "packing a suit inside a groan". The Nonword stimuli were a control condition, explained in more detail below.

ID	Stimulus	Mean rank
a	swoolding a dasck made of a boarfth	5.52
b	drylping a spraup made of a smoave	5.90
c	whupting a gleife using a traufe	6.64
d	scaphthing a jourfth with a squiep	6.68
e	balning a strolgn with a trofth	6.70
f	squoaping a prulv with a grenk	6.70
g	pholgning a plauvv using a whefth	6.90
h	groonthing a dwimf with a champt	7.04
i	phleabing a smanse made of a squald	7.04
j	gognthing a trylse using a shrorfe	7.20
k	hilcing a gnafth with a drolce	7.32
l	wuipthing a yuifth using a norfth	7.40
m	julning a thwirze made of a thweuff	7.64
n	saffthing a fenks using a hurgnth	8.10
o	twimbthing a toamth made of a thrirc	8.22

Table 3: Stimuli in Nonword control condition of Exp 1, sorted from most likely (top) to least likely (bottom).

Since our primary question was whether people make graded judgments *within* the categories of impossible or inconceivable event descriptions, we first wanted to ensure that these items were robustly considered to be impossible or inconceivable. To do this, we used the data from Hu et al.'s categorization experiment, where participants were asked to categorize each event description as "probable", "improbable", "impossible", or "nonsense". We filtered the stimuli used in those studies, keeping only items where at least 75% of participants agreed on the expected label ("impossible" for the Impossible condition, and "nonsense" for the Inconceiv-



(a) (b)

Figure 1: (a) Mean ranks measured in Experiment 1 for each item, across three stimulus sets and simulated random ranking data. (b) Mean Pearson r correlation between the mean rankings from randomly sampled halves of participants.

able condition). Of these, we selected 15 from those subsets for the Impossible and Inconceivable conditions.

As a negative control, for our third stimulus set we constructed 15 strings that follow the same syntactic structure as the Impossible and Inconceivable conditions, but with content words replaced with English nonwords: for example, "gogn-thing a trylse using a shrorfe". A priori, we would *not* expect to see systematic agreement between people for these non-words. Finding such an agreement would indicate that the gradedness task is not interpretable for the Inconceivable condition (and would also call into question the interpretation of gradedness for the Impossible condition). To construct these control stimuli, we first created 45 nonwords from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002). We only selected nonwords with orthographically existing onsets, legal bigrams, and 0 phonological neighbors. We also restricted nonwords to be between 3-7 letters. After generating this set of nonwords, we randomly arranged them into 15 groups of 3, where for each group the 3 nonwords correspond to the verb, the object, and a tool or prepositional object. We then "conjugated" the nonword in the verb position to follow English morphology. In order to control for potential effects of the tool or relationship to the verb, we also varied the tool description: 5 of the 15 items followed the object with "with a"; another 5 followed the object with "using a"; and the remaining 5 followed the object with "made of a".

Data collection For each of the three stimulus sets, we recruited participants on Prolific, based in the USA, with a self-reported native language of English. Participants were asked to rank 15 sentences by dragging and dropping them in order, given the following instructions: *"Drag and drop the following sentences according to how likely they are to happen. Remember, a HIGHER-ranked sentence is MORE likely: e.g., #1 is the most likely, and #15 is the least likely. Please interpret the sentences in a LITERAL way, based on what is possible given the physical laws of the real world."*

We recruited 150 subjects for the Impossible and Inconceivable stimuli, with one participant dropped due to data corruption. Participants were randomly assigned to one of the two conditions, resulting in 83 subjects for Impossible and 66 subjects for Inconceivable. We recruited 50 subjects for the Nonword stimuli. Participants were paid $12/hour.

Results Figure 1a shows the mean ranking for each item across all stimulus sets. To test whether the rankings were different from uniform, we performed Friedman's two-way analysis of variance by ranks. We found that the rankings for the Impossible and Inconceivable conditions were significantly different from the uniform baseline, suggesting that there was systematic agreement between people on their degree of gradedness (Impossible: Friedman $F = 210.09$; Inconceivable: $F = 104.24$; all conditions have a critical value of $\chi^2(df = 14, p = 0.05) = 23.68$).

As an additional control, we also generated random ranking data by simulating N uniformly random rankings of 15 items, where N is the number of participants in the Inconceivable ranking study. In the two control conditions, the Nonword rankings and the Random simulated rankings were *not* significantly different from a uniform ranking (Nonword: $F = 18.09$; Random: $F = 9.01$). This suggests the results in the Impossible and the Inconceivable are not simply the result of how the procedure of sorting ranking works, nor general agreement to any language-like collection of tokens.

As another measure of agreement, for each dataset we computed the Pearson correlation between mean rankings, for randomly sampled halves of the participants. Figure 1b shows the resulting correlation coefficients, averaging over 100 randomly sampled halves. The average correlation is strongest for the Impossible (r =0.91) and Inconceivable (r =0.79) conditions, and much weaker for the nonword (r =0.16) and random ($r = -0.18$) conditions. Again, this analysis confirms that there appears to be structured agreement within impossible and inconceivable event descriptions.

Our results for the Impossible stimuli reproduce the gradedness pattern found by prior studies of "shades of impossibility". Our results for the Inconceivable stimuli provide the first evidence of structured agreement in some inconceivable things being more likely than others. Given this finding, in our next two experiments, we consider what leads people to

think some nonsense makes more sense.

Experiment 2: Does metaphorizability explain graded inconceivability?

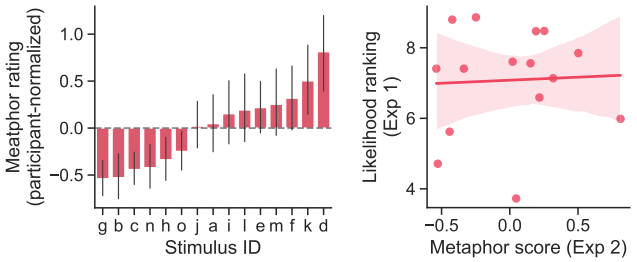
In Experiment 1, we explicitly instructed participants to interpret sentences literally. But, it is possible that they (intentionally or unintentionally) attributed figurative meanings to the sentences, and then used these meanings to rank the sentences. For example, "building a house out of language" is not literally possible, since language is not the kind of thing that rigid objects can be built out of, but it may be that people interpreted it to metaphorically mean something like using language to create a shared sense of community. We note that for our purposes it does not matter whether people actually *had* a specific interpretation in mind—it was sufficient that an item merely seemed like the kind of thing that one could find a metaphorical interpretation for. Returning to the previous example, it may be that "building a house out of language" seems like a good metaphor, even if one does not come up with an explanation for what that metaphor *is*.

In Experiment 2, we tested the notion that the graded likelihood of inconceivable items is based on metaphorizability, by examining whether how "good" of a metaphor an inconceivable sentence is predicts how likely it is perceived to be.

Data collection We used the same 15 inconceivable event descriptions as in Experiment 1 (Table 2). At the beginning of the experiment, participants were told "We're trying to come up with new metaphors, and we need your help." On each trial, they were asked "Can you rate how good of a metaphor this sentence is?" and then presented with the event description (e.g., "growing flowers inside a sneeze"). They provided their response on a slider with endpoints labeled "Very bad metaphor" (internally coded as 0) and "Very good metaphor" (internally coded as 100). After providing the rating, participants saw a screen saying "If you thought the previous metaphor was a good one, can you explain what it means? If not, you can respond with 'Not sure' and continue to the next trial." We recruited 36 participants on Prolific, with US-based IP addresses and a self-reported native language of English.

Results Figure 2a shows the mean metaphor rating for each of the 15 event descriptions, with ratings normalized within-participant. We found that there is structured agreement across people: across 100 randomly sampled halves of participants, the mean correlation between ratings was 0.76. However, these judgments of metaphor quality do *not* predict the mean rankings of "likelihood" obtained in Experiment 1, as shown in Figure 2b (Pearson $r = 0.05$, $p = 0.87$).

The results from Experiment 2 suggest that people's judgments of the likelihood of inconceivable events are not explained by appealing to figurative language or non-literal interpretations. Even though these interpretations exist, and seem to be agreed upon by people, being a "better" metaphor does not necessarily make an inconceivable event more likely.



(a) (b)

Figure 2: (a) Mean "metaphor quality" score (normalized within-participant) for each stimulus. Stimulus IDs correspond to Table 2. (b) Metaphor score (x-axis) does not predict mean likelihood ranking from Experiment 1 (y-axis).

If non-literal interpretation is not the explanation, then what might be driving systematic gradedness in inconceivability? An alternate explanation is that people are trying to make sense of nonsense in a literal way, but hitting a specific kind of mental "roadblock", equivalent to a kind of type error. Moreover, some roadblocks are worse than others, leading to the structure observed in the rankings. We test one specific version of this hypothesis in Experiment 3.

Experiment 3: Do type distances explain graded inconceivability?

Here, we return to the idea of inconceivability as analogous to the mind hitting a "type error" (e.g., Sosa & Ullman, 2022; Hu et al., 2025; Magidor, 2017). On this view, inconceivability refers to a category of things that cannot be evaluated or computed, due to being outside the scope of what is able to be thought within a given mental module. Computationally, this scope is defined by a type system that dictates what computational primitives are available, and the available rules for composing those primitives. The stimuli being evaluated are like inputs to a program, and what makes an event description nonsensical is whether it disobeys certain type-constraints, like trying to input `x='armadillo'` into a `sqrt(x)` program. This view builds on foundational work in the computational theory of mind (Feigenbaum, Feldman, et al., 1963; Fodor, 1975; Pylyshyn, 1980) and developmental psychology (Keil, 1979; Sommers, 1971), as well as more recent work that models aspects of cognition as the running, modifying, or learning of different programs in a domain-specific language (Wong et al., 2023; Griffiths, Chater, & Tenenbaum, 2024; Gerstenberg & Tenenbaum, 2017; Ullman & Tenenbaum, 2020).

In Experiment 3, we tested the idea that rankings of type violations depend on the "severity" of the type violation. Specifically, violating the expected type with a similar type is less of an error than violating it with a dissimilar type.

Stimuli We created a novel set of stimuli grounded in a hierarchy of basic types (Figure 3), inspired by previ-

Condition	Function	Target type	Term	Stimulus
Target ●	$\lambda x : x$ feeling regretful	Concrete-Animate	a man	a man feeling regretful
Target ●	$\lambda x : x$ being entertained by a podcast	Concrete-Animate	a boy	a boy being entertained by a podcast
Target ●	$\lambda x :$ adding x to 79	Abstract-Numeral	the number 20	adding the number 20 to 79
Near ●	$\lambda x : x$ being greater than 10	Abstract-Numeral	an inch	an inch being greater than 10
Near ●	$\lambda x :$ dividing 12 by x	Abstract-Numeral	a mile	dividing 12 by a mile
Near ●	$\lambda x : x$ being surprised by a news article	Concrete-Animate	a shirt	a shirt being surprised by a news article
Far1 ●	$\lambda x : x$ being a prime number	Abstract-Numeral	a girl	a girl being a prime number
Far1 ●	$\lambda x :$ taking the square root of x	Abstract-Numeral	a toddler	taking the square root of a toddler
Far1 ●	$\lambda x : x$ writing a play	Concrete-Animate	the number 13	the number 13 writing a play
Far2 ●	$\lambda x : x$ being amused by a TV show	Concrete-Animate	a month	a month being amused by a TV show
Far2 ●	$\lambda x : x$ applying for a job	Concrete-Animate	an inch	an inch applying for a job
Far2 ●	$\lambda x : x$ being a multiple of 5	Abstract-Numeral	a door	a door being a multiple of 5

Table 4: Stimuli used for ranking in Experiment 3.

ous work in developmental psychology on how children might represent ontologies of the world as tree-like structures, where nodes are objects and edges are relations between objects (Sommers, 1971; Keil, 1979; Schmidt, Kemp, & Tenenbaum, 2006). The relationships between types are represented by a two-level binary tree, with top-level categories CONCRETE and ABSTRACT, and sub-levels ANIMATE/INANIMATE (within CONCRETE) and NUMERAL/QUANTITY (within ABSTRACT). The "distance" between types can then be viewed as distance in tree traversal: going from ANIMATE to INANIMATE only requires moving along 2 edges, whereas going from ANIMATE to NUMERAL or QUANTITY requires moving along 4 edges.

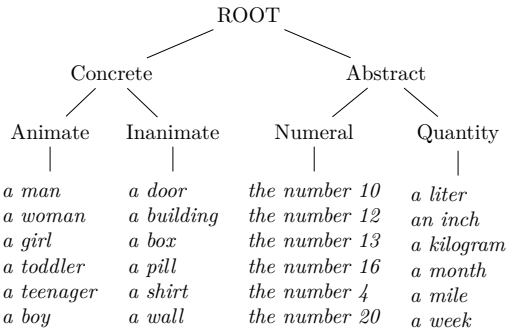


Figure 3: Hierarchy of types used to create stimuli for Exp 3.

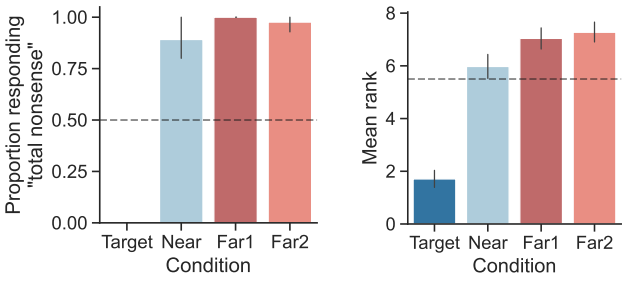
We designed a set of 12 functions: 6 that take ANIMATE as the expected type (e.g., `x feeling regretful`), and 6 that take NUMERAL as the expected type (e.g., `taking the square root of x`). We then constructed stimuli by crossing the 12 functions with items from the 4 types from our hierarchy. In the Target condition, the function is applied to an item with the expected type (e.g., "a man feeling regretful"). In the Near condition, the function is applied to an item that is close to the expected type in our hierarchy (e.g., "a door feeling regretful"). Finally, in the Far conditions, the function is applied to an item that is far to the expected type in our hierarchy. Since there are two types that are "far" (i.e., 4 edges away) from each type, we have two instances of the Far condition for each function (e.g., "the number 10 feeling regretful" and "a liter feeling regretful").

Norming & stimulus selection Before conducting the main experiment, we wanted to confirm that all "type errors" in our stimuli were perceived as type errors, regardless of whether the violating type was near or far from the expected type. We conducted a norming study asking participants to categorize each statement as "Total nonsense" or "Not total nonsense". We recruited 40 US-based participants on Prolific, and each participant categorized 12 statements.

As expected, participants rarely categorized the statements in the Target condition as "total nonsense" (4%), and nearly always categorized the statements in the Far conditions as "total nonsense" (84% for Far1, 89% for Far2). Importantly, participants categorized statements in the Near condition as "total nonsense" well above chance (68%). This confirms that on the whole, statements in the Near condition are akin to a type violation. However, there was still a difference between for Near and Far (Near vs. Far1: $t = -3.06$, $p = 0.002$; Near vs. Far2: $t = -4.20$, $p < 0.001$), which we take to suggest that people are able to find valid interpretations of poorly-typed sentences. We return to this point in the Discussion.

We used the norming data to select stimuli for the main ranking experiment. We took the items from the Target condition with the lowest rate of "total nonsense" responses, and the items from all other conditions with the highest rate of "total nonsense" responses. We combined these into a list of 12 items, such that each of the 12 functions is seen exactly once, and at least one instance of each type (ANIMATE, INANIMATE, NUMERAL, and QUANTITY) is seen in each condition. This normed list of 12 items was then used for the main ranking experiment.

The proportion of trials where people labeled items as "total nonsense" for the final, normed stimuli is shown in Figure 4a. Importantly, for these items, the proportion of "nonsense" responses was not significantly different between the Near and Far conditions (Near vs. Far1: $t = -1.86$, $p = 0.14$; Near vs. Far2: $t = -1.34$, $p = 0.25$), suggesting that the vi-



(a) (b)

Figure 4: (a) Results from norming study: Proportion of trials where people labeled the statement as "total nonsense" for the final set of 12 normed items. (b) Results from main study (ranking): mean ranking of statements in each condition.

olations involving types near the expected type were indeed perceived as type errors.

Data collection Before the main task, participants completed a brief multiple-choice comprehension check ("What is your task in this study?"). During the main ranking task, participants saw the same interface and instructions as in Experiment 1. We recruited 60 US-based participants on Prolific with a native language of English. 4 participants were dropped due to failing the comprehension check or data corruption, resulting in data from 56 participants.

Results Figure 4b shows the mean ranking for each condition. Items in the Target condition are judged as the most likely (i.e., lowest mean ranking) overall, and items in the two Far conditions are judged as the least likely (i.e., highest mean ranking). Importantly, we also observe that rankings in the Near condition are between the Target and Far conditions. The mean ranking in the Near condition is significantly different from the mean ranking in both Far conditions (Near vs. Far1: $t = -3.45$, $p < 0.001$; Near vs. Far2: $t = -4.24$, $p < 0.001$), and the mean rankings in the Far1 and Far2 conditions do not differ from each other ($t = -0.81$, $p = 0.42$).

As mentioned previously, these rankings are correlated with the proportion of "total nonsense" labels from the norming on our stimuli. Assuming that people are doing something akin to a type-based inference, these results suggest that when evaluating inconceivable statements, people rely on a hierarchy of types such that "worse" type violations are seen as more nonsensical. Further, it suggests that people are performing "type coercion", which we discuss in detail below.

Discussion

Graded impossibility has been a topic of focused research in cognitive science and philosophy, for good reason. But research also suggests that the *impossible* is different from the *inconceivable*. We examined whether inconceivable, nonsensical, category-error events will also be graded in their nonsense, and why. We found that people have systematic, graded judgments about nonsensical statements (Experiment 1), and then tested two potential explanations for why (Experiments 2 and 3). Experiment 2 showed metaphorizability is systematic, but doesn't correlate with likelihood, suggesting that while there is structure in how people form metaphors, it is independent of degrees of inconceivability. In Experiment 3, we turned to an alternative theory based on the notion that the boundaries between conceivable and inconceivable are defined by a mental type system. We developed a set of stimuli on the basis of a simple hierarchy of types such as CONCRETE/ABSTRACT, and observed that the distance between types in this hierarchy explained the rankings given by people. We emphasize that the ontology of types we used is not meant as a strong statement about the specific cognitive ontology people may use, but as a first-pass approximation of the general theory that people use a hierarchy of types to make graded sense of inconceivability.

Even given a hierarchy of types, one needs to explain how to connect type violations to a graded judgment of inconceivability. After all, there are (existing) programming languages in which all type errors are themselves of the same type. Whether trying to perform the computation `addition(2, 'armadillo')` (a violation of an expected *float* type with a *string* type), or the computation `flip(5)` (a violation of the expected *list* type with an *int* type), one may encounter a similar breakdown in computation resulting in `TypeError` being returned. Such a process does not explain gradedness. But, we can help ourselves to another basic computational process known as "type coercion". In type coercion, a compiler can force the type of an entity, to make sense of a type error. The process works by first encountering a type error, and then attempting to re-interpret the input or process along specified lines, to produce valid output. For example, `addition(2, 'armadillo')` may trigger an initial type error. But, rather than fail, the compiler would interpret `2` as a *string* rather than an integer, and `addition` as string concatenation rather than mathematical addition, finally returning `'2armadillo'`.

The suggestion is then: (1) people use something akin to type-coercion to handle type errors (inconceivable statements); (2) it is easier to coerce a type that is closer in a type hierarchy ("easier" could apply to different parts of the computational process—e.g., a coerced output takes less time to produce, or a function is more likely to return an output); (3) this ease is graded; and (4) is the basis for the graded judgments of nonsense. This suggestion is not the end-point explanation, but a mid-point that raises further questions, including how cognitive type coercion works more specifically, and what part of the process is used to estimate gradedness.

Some theories are more wrong than others. But, in a phrase often attributed to Pauli, some theories are "not even wrong". Yet it seems that some things are not-even-wrong-er than others. The right cognitive theory of such judgments of wrongness can help us understand how people reason about the everyday, more reasonable world. Does that make sense?

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.

References

Byrne, R. M. (2007). *The rational imagination: How people create alternatives to reality*. MIT press.

Feigenbaum, E. A., Feldman, J., et al. (1963). *Computers and thought* (Vol. 37). New York McGraw-Hill.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Gendler, T. S., & Hawthorne, J. (2002). *Conceivability and possibility*. Clarendon Press.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. R. Waldmann (Ed.), *The oxford handbook of causal reasoning* (pp. 515–547). Oxford, UK: Oxford University Press.

Griffiths, T. L. (2015, March). Revealing ontological commitments by magic. *Cognition*, *136*, 43–48. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010027714002212 doi: 10.1016/j.cognition.2014.10.019

Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (2024). *Bayesian models of cognition: reverse engineering the mind*. MIT Press.

Hu, J., Sosa, F., & Ullman, T. (2025). Shades of zero: Distinguishing impossibility from inconceivability. *Journal of Memory and Language*, *143*, 104640. Retrieved from https://www.sciencedirect.com/science/article/pii/S0749596X25000336 doi: 10.1016/j.jml.2025.104640

Keil, F. C. (1979). *Semantic and Conceptual Development*. Harvard University Press. Retrieved 2024-01-30, from https://doi.org/10.4159/harvard.9780674181816 doi: 10.4159/harvard.9780674181816

Lewis, D. K. (1986). *On the plurality of worlds*. Oxford: Blackwell.

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

Magidor, O. (2009, December). Category mistakes are meaningful. *Linguistics and Philosophy*, *32*(6), 553–581. Retrieved from https://doi.org/10.1007/s10988-010-9067-0 doi: 10.1007/s10988-010-9067-0

Magidor, O. (2017, January). Category mistakes and figurative language. *Philosophical Studies*, *174*(1), 65–78. Retrieved from https://doi.org/10.1007/s11098-015-0575-1 doi: 10.1007/s11098-015-0575-1

McCoy, J., & Ullman, T. (2019, May). Judgments of effort for magical violations of intuitive physics. *PLOS ONE*, *14*(5), e0217513. Retrieved from https://doi.org/10.1371/journal.pone.0217513 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0217513

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain sciences*, *3*(1), 111–132.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology*, *55A*, 1339–1362.

Schmidt, L. A., Kemp, C., & Tenenbaum, J. B. (2006). Nonsense and Sensibility: Inferring Unseen Possibilities. In *Proceedings of the Annual Meeting of the Cognitive Science Society.* Retrieved from https://escholarship.org/uc/item/7t32n7z0

Shtulman, A., & Morgan, C. (2017, October). The explanatory structure of unexplainable events: Causal constraints on magical reasoning. *Psychonomic Bulletin & Review*, *24*(5), 1573–1585. Retrieved from https://doi.org/10.3758/s13423-016-1206-3 doi: 10.3758/s13423-016-1206-3

Sommers, F. (1971, January). Structural ontology. *Philosophia*, *1*(1), 21–42. Retrieved from https://doi.org/10.1007/BF02378925 doi: 10.1007/BF02378925

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.* Retrieved from https://arxiv.org/abs/2210.01634

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*(1), 533–558.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.