Decomposing modal thought

Jonathan Phillips[1,3] and Angelika Kratzer[2,3]

[1]Dartmouth College, Hanover NH

[2]University of Massachusetts Amherst, MA

[3]Centre for Advanced Studies in the Humanities 'Human Abilities', Berlin

Abstract

Cognitive scientists have become increasingly interested in understanding how natural minds represent and reason about possible ways the world could be. However, there is currently little agreement on how to understand this remarkable capacity for modal thought. We argue that the capacity for modal thought is built from a set of relatively simple component parts, centrally involving an ability to consider possible extensions of a part of the actual world. Natural minds can productively combine this ability with a range of other capacities, eventually allowing for the observed suite of increasingly more sophisticated ways of modal reasoning. We demonstrate how our (de)compositional account is supported by both the trajectory of children's developing capacity for reasoning about possible ways the world could be and by what we know about how such modal thought is expressed within and across natural languages. Our approach makes new predictions about which kinds of capacities are required by which kinds of experimental tasks and, as a result, contributes to settling currently open theoretical questions about the development of modal thought and the acquisition of modal vocabulary in children. Our work also provides a more systematic way of understanding possible variation in modal thought and talk, and, more generally, paves the way towards a unified theory that will ultimately allow

researchers across disciplines to relate their findings to each other within a framework of shared assumptions.

Keywords: modality; anchor semantics; counterfactual reasoning; actuality-directed modal cognition; metacognition

Author note. These studies were approved by Dartmouth College's Committee for the Protection of Human Subjects (STUDY00032209). All experimental details, stimuli, materials, data, code, and a longer explication of the results can be found in the supplement to our paper: https://doi.org/10.7910/DVN/KUWNYK. Previous version of these ideas were presented at the annual meeting of the Society for Philosophy and Psychology, the New York Philosophy of Language Workshop, the Harvard Language & Cognition talk series, the MIT Philosophy and Linguistics Workshop, the Centre for Advanced Studies in the Humanities, "Human Abilities" in Berlin, the University of Göttingen Psychology Department, the Central European University Cognitive Development Seminar, the Semantics Workshop at UMass Amherst, and the Experiments in Linguistic Meaning conference. A preprint of this paper was made available on PsyArXiv at: https://psyarxiv.com/g6tzc/.

Acknowledgments. We'd like to thank the institutions that made our collaboration possible: The 2015/2016 SIAS (Some Institutes of Advanced Study) Summer Institute on the Investigation of Linguistic Meaning, funded by the Alexander von Humboldt Foundation and the Andrew W. Mellon Foundation and hosted by the Wissenschaftskolleg Berlin and The National Humanities Center, The Radcliffe Institute for Advanced Study of Harvard University, The Centre for Advanced Studies in the Humanities 'Human Abilities' in Berlin, and the Diesel Café in Somerville. For detailed comments, discussions, and other forms of support we thank Fiery Cushman, Mariel Goddu, Thomas Icard, Joshua Knobe, Daniel Lassiter, Matt Mandelkern, Brian

Leahy, Barbara Vetter, Michael Waldmann, and three very helpful anonymous reviewers for The Psychological Review.

Decomposing modal thought

Across the Cognitive Sciences, there's a newfound interest in studying modal cognition, our ability to represent possible ways the world could be—whether it be representing what could happen in the future, reasoning about what might have happened in the past, figuring out what could have happened if things had been different, or even just guessing what else might be the case right now in parts of the world beyond our reach. This kind of modal thought is actuality-directed and is central to much of high-level human thought, including for example, causal and counterfactual reasoning (Lewis,1973, 1974; Pearl, 2009; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021), or planning and decision making (Kaelbling, Littman, & Cassandra, 1998; Morris, Phillips, Huang, Cushman, 2021), or judgments of moral responsibility (Acierno, Mischel, & Phillips, 2022; Phillips & Knobe, 2008; Woolfolk, Doris, & Darley, 2006).

Historically, work under the label of 'modality' has been the pursuit of philosophers, linguists, logicians, or computer scientists. But these days, an increasing number of researchers are studying how natural minds actually represent and reason about possible ways the world may be or could have been—from non-human animals (e.g., Engelmann, Völter, O'Madagain, Proft, Haun, Rakoczy, & Herrmann, 2021; Engelmann, Völter, Goddu, Call, Rakoczy, & Herrmann, 2023; Redshaw, & Suddendorf, 2016), to human infants (e.g., Cesana-Arlotti, Téglás, & Bonatti, 2012; Téglás & Bonatti, 2016; Téglás, Girotto, Gonzalez, & Bonatti, 2007) and young children (e.g., Alderete & Xu, 2023; Goddu, Sullivan, & Walker, 2021; Leahy, Huemer, Steele, Alderete, & Carey, 2022; McCormack & Hoerl, 2020; Redshaw, Suddendorf, Neldner, Wilks, Tomaselli, Mushin, & Nielsen, 2019; Robinson, Rowley, Beck, Carroll, & Apperly, 2006; Shtulman & Carey, 2007), to adult humans (e.g., Byrne, 2007; De Brigard, Addis, Ford, Schacter, & Giovanello, 2013; Morris, Phillips, Huang, & Cushman, 2021; Phillips & Cushman, 2017).

In this paper, we will be exclusively concerned with actuality-directed modal thought. Actuality-directed modal thought aims at figuring out what the actual world is like by considering and evaluating possible extensions of an accessible part of it. It thus literally involves reference to a part of the actual world. Here is an illustration. Suppose you are asking a friend whether she happens to have a cardboard box you could use for shipping a cast iron pot to your brother. She points to a box sitting on one of her shelves and says: "You can ship the pot in this box." However, unbeknownst to her and unfortunately for you, the bottom of the box had gotten wet and started to decompose. It couldn't possibly hold a cast iron pot. Given all of this, it turns out that what your friend told you was false. Whether you can ship your pot in the box she pointed at thus depends on the actual state of the box, and in this case, this is so regardless of what she may know about it. As this example illustrates, actuality-directed modal thought involves taking some part of the actual world–the modal anchor in the terminology of Hacquard (2006)–and then, in this case, involves entertaining possible *future* extensions of it. In our example, the modal anchor is the actual situation we found ourselves in, which crucially includes the actual box with all of its flaws and imperfections, along with the pot to-be-shipped. As a consequence, all future possibilities that evolve from the anchor situation will have a version of exactly that box with those same flaws and imperfections in it. And so if things develop normally, there aren't any future possibilities where the pot is successfully shipped to your brother. Modal thinking of this kind is actuality-directed in our sense because it is facts of a part of the actual world that (together with normality assumptions) determine what is or isn't possible.

In its most general form, thought about possibilities is pervasive, but, crucially, not all thought that is concerned with possibilities is actuality-directed in the sense we are targeting here. For example, pigeons can famously be trained to associate pecking at a target with the subsequent production of a food reward, and with enough experience, can further learn that this association only holds under certain conditions and not others (Skinner, 1965). This kind of reinforcement

learning involves pairing an action with as-of-yet non-actual events—involving the future emergence of food—and thus can be understood as a capacity that concerns future possibilities. Yet simply coming to learn associations through repeated exposure is not an instance of actuality-directed thought of the kind we are interested in here. Or, to take another example, consider a sentence like *there is a cat on my lap*. You know its meaning if you know which possible situations would make it true: not just some actual situation you may be talking about that happens to have a cat curled up in your lap, but any possible situation that has a cat on your lap. Understanding the meaning of even the simplest sentence, then, relies on our capacity to connect sentences to possibilities. Yet merely understanding sentences is again not an instance of actuality-directed modal thought of the kind our account is meant to cover. In a similar vein, pretense play requires the ability to imagine mere possibilities and convey and understand information about them (Harris, 2022). Yet this kind of thinking about possibilities is again not actuality-directed in our sense. Like fiction, pretense play may stay close to the actual world in certain ways, but it is not necessarily anchored in actual situations and is not necessarily concerned with figuring out what the actual world may look like beyond the limits of the anchor situation. We will set thought about imaginary possibilities aside for now but will return to this question at the end of the following section when comparing our account to other approaches to modal thought and to thought about possibilities more generally.

In the growing body of empirical work on modal cognition in comparative, developmental, and cognitive psychology, there has been remarkably little agreement on how modal thought works, or what exactly it involves. One of our aims in this paper is to contribute to clarity about these questions. Unifying insights across linguistics, philosophy, logic, and computer science, we argue that what has been discussed under the label ‘modal cognition’ in the recent developmental literature, for example, is largely actuality-directed modal thought in our sense and is built from a set of relatively simple component abilities, all involving the basic ability of

considering possible extensions of an actual anchor. This capacity for considering possible extensions of an actual anchor situation is thus a common component of many apparently different types of modal cognition. It can be refined further or productively combined with other capacities, yielding more sophisticated abilities such as the capacity to consider possibilities that were still live options at some point in the past, but that we now know to be only counterfactually possible, the capacity to compare, rank, weight or quantify over possibilities, or the capacity to design an optimal plan of action in light of a range of possibilities. It is these more complex capacities that ultimately give rise to the sophisticated range of modal reasoning found in adult humans.

The theoretical contribution of our work is threefold. First, our (de)compositional approach makes it possible to relate the results of existing developmental studies to each other and to use them towards a new way of theorizing about variation in modal thought—whether it be variation across species, within human development, or throughout natural languages. Second, our approach makes new predictions about which experimental tasks require which kinds of psychological capacities. As a result, we are able to settle open questions and current disagreements about the trajectory of the development of modal thought and the acquisition of modal vocabulary in children. Third, our work offers a foundation for the integration of empirical and theoretical insights across disciplines; it shows, for example, how the discoveries in comparative psychology fit with the data from natural language semantics, or how we can unify the representational proposals from developmental and cognitive psychology with the theoretical frameworks made at the level of formal logic or computational models.

Components of modal thought: illustrations

To get a first sense for the ingredients of actuality-directed modal thought, let's consider a simple case in more detail. Varying our earlier cardboard box example, imagine that you are deliberating on what to pack in a particular box in view of an upcoming move. Such an ability can be decomposed into a set of component processes (see Fig 1). The first step consists in choosing the modal anchor. The modal anchor is the part of the actual world that contains all the facts that are relevant for the possibilities we are considering. Since you are wondering what to pack in a particular box, you would take as your anchor the actual situation that includes the empty box, whatever objects are waiting to be packed, and any other relevant parts of the situation you are in. In a second step, you consider possible four-dimensional extensions of this anchor situation: possible situations that take off from exact matches of the anchor but evolve into possibly divergent futures from there. Those possible futures might differ as to which things go into the box. You might, for example consider a future extension of the anchor situation where books are put in the box, or one where it is a ceramic pot, or perhaps one where it is a pair of shoes.

The process of considering possible extensions of an actual modal anchor can be understood as implementing an ability for *factual domain projection*, that is, for a process that implements a function, f_{act}, which takes a part of the actual world–the anchor–as its argument and returns a set of possible situations containing exact matches of the anchor (Kratzer, 2013).[1] At the most abstract, mathematical, level, f_{act} would be a function that relies on a model structure consisting

[1] Of course, there are many ways that such a function may be realized at an algorithmic level, and even more at the level of physical implementation; our proposal is not an account at these levels of analysis (Marr, 1982). We will return to the connection between our proposal and these other levels of analysis at the end of the paper.

of a set of possible situations and two binary relations on that set: a match relation and a part relation where worlds are maximal situations. Towards the end of the paper we will have more to say about possible implementations of f_{act}, as well as about its crucial role in providing a bridge that connects the different disciplines that investigate modal thought at different theoretical levels. In the meantime, we will mostly talk about f_{act} in an informal or semi-formal way.

The result of factual domain projection is a set of possible situations bound to the actual world by matches of the anchor. Collectively, these possibilities make up the *modal domain*, which one could go on to reason about by comparing, ranking, weighting, or quantifying over the possibilities in it. In our example, you might go on to reason that, while you could pack books in that box, the best thing to pack would be the ceramic pot.

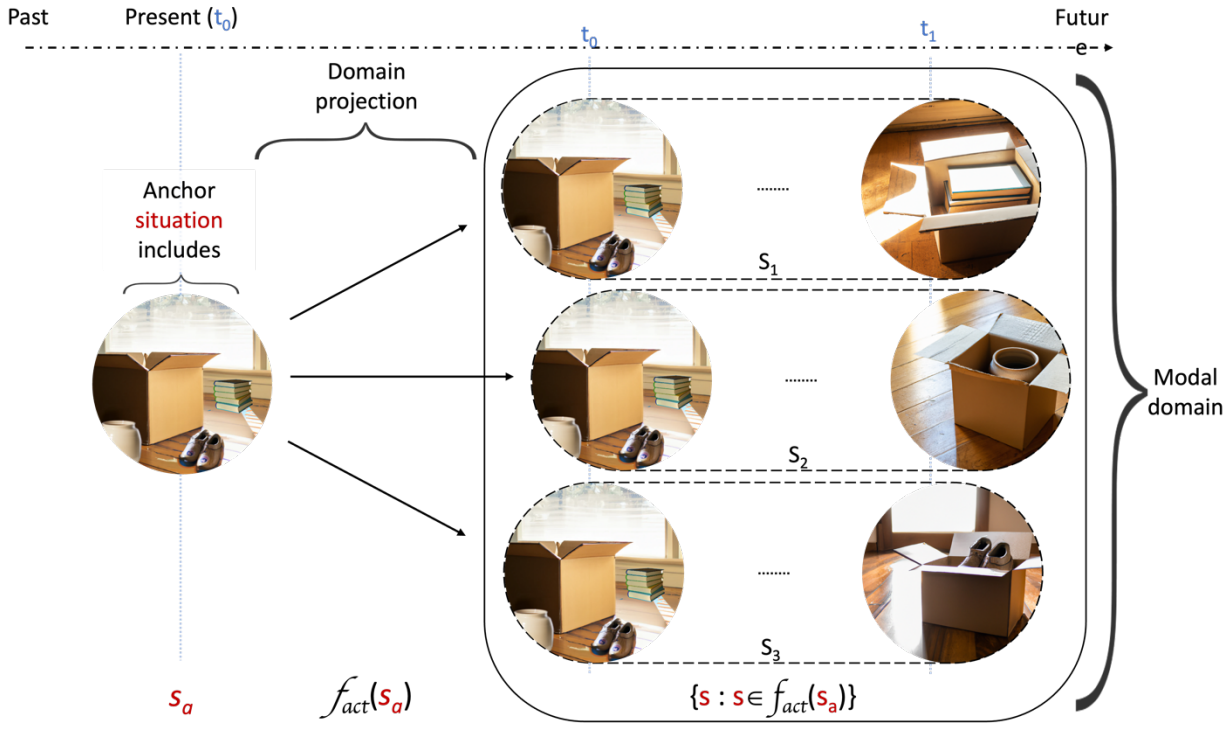


Figure 1. Schematic depiction of the core processes of future-oriented modal thought when considering what to pack inside a box. A function f_{act} instantiates an inverse projection, taking as input the modal anchor s_a (which is an actual situation at t_0 that includes the box you are packing) and returning a set of situations, $\{s: s \in f_{act}(s_a)\}$, with matches of the anchor at the present time but possibly different extensions into the future. Among these are situations in which you pack books, a ceramic pot, or shoes in the box, s_1-s_3, respectively.[2]

To have another example, imagine now that you are looking at a puddle of water on the ground and are wondering where it came from (Fig. 2).[3] The process you use is much the same as in the case with the box: you take as your anchor an actual situation that includes the facts that are relevant for the possibilities you are considering and then generate possible past extensions of that situation. In this case, your modal anchor would include the actual puddle, and possible

[2] The images in these schematic depictions of factual domain projection were all generated by JP using the new outpainting feature in DALL·E 2. It was a fortuitous coincidence that OpenAI developed their artificial intelligence tool for generating extensions of an input image at the same time we were developing our theory of how natural intelligence generates extensions of an anchor situation.

[3] Some may find it helpful to note the similarity between the kind of modal inferences we are discussing and the kinds of inferences that are often required in processing perceptual information. They are both forms of 'inverse projection problems.' In the case of vision, for example, one has to infer the 3-dimensional structure of the world from 2-dimensional images projected onto one's retinas. There are many possible ways the 3-dimensional world could be such that it would project *those* two-dimensional images, and one has to reason over these possibilities to infer which 3-dimensional world is actual. Modal inferences have a similar structure: one has to infer some higher-dimensional representation of the actual world from some observed lower-dimensional piece of that world. There are, again, many possible ways that the actual world may be such that one would observe this particular piece of the world to be the way that it is, and one must reason over that set of possibilities to infer which higher-dimensional state of the world is actual.

extensions of the anchor would include possible earlier events leading to that puddle: a storm passing through, a bucket of water being spilled, or a fire hydrant being tested.

The puddle example differs in an important way from the box example. While the anchor is a part of the actual world at the present time in both, the possible extensions considered stretch forward into the future in the case of the box, but backwards into the past in the case of the puddle. That is, the two examples illustrate a change in the *temporal orientation* of the extensions of the anchor situation that are being considered (Condoravdi, 2002). The basic ability to consider possible extensions of an actual anchor situation can thus be refined by distinguishing extensions according to their temporal orientation. They may stretch forward into the future or backwards into the past, and they could also stay within the present, with matches of the anchor situation appearing in varying possible spatial surroundings.

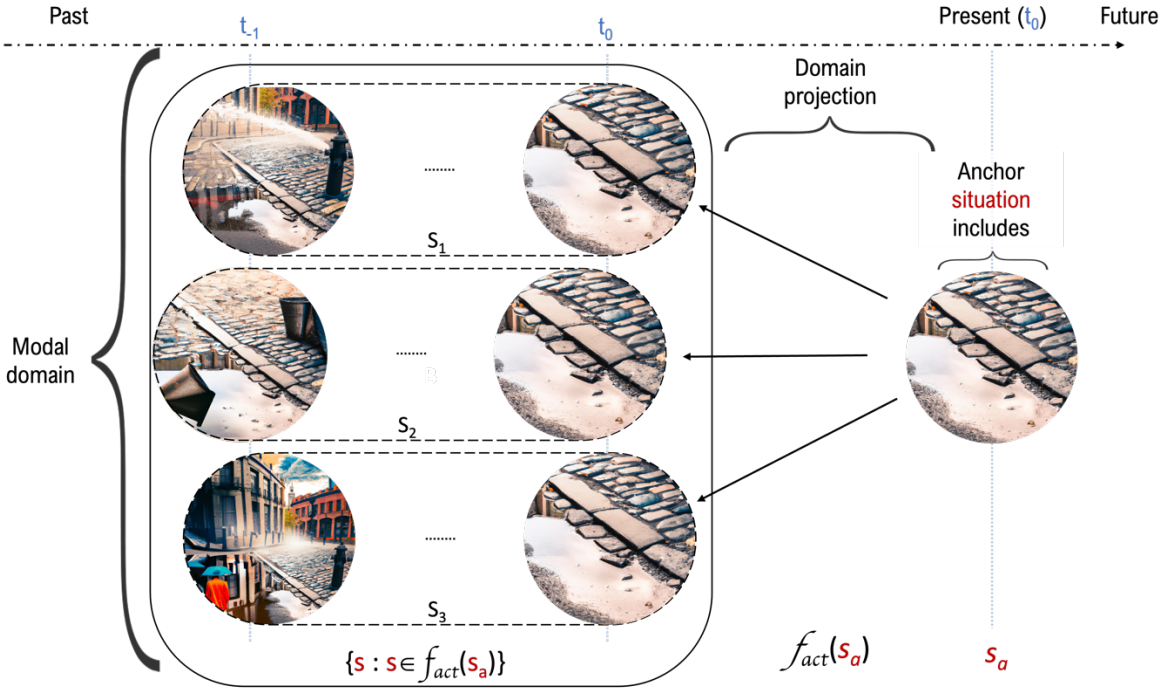


Figure 2. Schematic representation of the core processes of modal cognition when reasoning about the causes of a puddle on the ground. A function f_{act} instantiates an inverse projection,

taking as input the modal anchor s_a (an actual current situation involving the puddle you are looking at) and returning a set of past situations, $\{s: s \in f_{act}(s_a)\}$, which have matches of the anchor situation at the present time, but different past extensions. Among these are situations in which a fire hydrant, a bucket of water, or a rainstorm caused the puddle, s_1-s_3, respectively.

Finding evidence for the proposed components

The central question we now face is what kind of evidence there is for this way of thinking about the abilities involved in actuality-directed modal cognition. One way to get traction on this question is to point out that our account makes predictions about developmental sequences and cross-species variation. On our proposal, we expect to find that more complex manifestations of modal cognition are the result of combining the foundation of factual domain projection with other separate capacities. Many observed manifestations of modal thought should thus be analyzable as the result of a combinatorics of basic abilities, and components of these same combinatorics should be detectable in patterns of variation across phylogeny and ontogeny. For example, the process of factual domain projection would be expected to exhibit variation that corresponds to differences in the choice of modal anchors or to differences in the temporal orientation of projected possibilities. Likewise, when examining how modality is encoded in natural languages, we might expect to find evidence that the combinatorics of abilities we posit for actuality-directed modal thought are reflected in the way the meanings of sentences expressing modal thoughts are compositionally constructed. In the sections to follow, we will provide evidence for our proposed decomposition, first from cognitive development, and then from natural language. Along the way, we will also present the results of three empirical tests that illustrate some of the more novel predictions of our proposal.

The form of the argument we build in what follows is that if one looks across all of these different sources of data—from developmental and comparative psychology, to the structure of natural languages, to our own experimental results—one sees a clear set of patterns emerge in how modal thought emerges and varies. It is those patterns that our account is meant to capture. In line with this approach, we have tried to review as much of the available data as is feasible while doing our best not to dwell too long on any single piece of evidence and all the possible ways that that particular piece of evidence may be accounted for.

Evidence from Cognitive Development

Future-oriented domain projection

Initial evidence for the ability to consider possible futures for present actual anchors can be found as early as twelve months. In a series of experiments by Téglás and colleagues (2007), infants watched as different objects bounced around a circular container that had a single opening at the bottom. Three of the objects were yellow, and one was blue. After infants watched the objects move around the container, an occluder covered the container, and a single object exited from its bottom. Twelve-month-old infants exhibited longer looking times when a blue rather than yellow objected exited, providing some initial evidence that they were representing future possibilities with yellow objects exiting the container. Critically, in a second study, the three yellow objects were blocked from exiting the container by an additional horizontal wall, and only the blue object could physically exit the container (see Fig. 3a). In this case, infants' looking time indicated the opposite pattern: they were more surprised when a yellow, rather than a blue, object exited, suggesting that they were no longer considering future possibilities where the yellow objects exited (Téglás et al., 2007).

These two studies serve to illustrate what has now been found across a range of studies: at a surprisingly early age, human children are able to project future possibilities from present anchor situations (e.g., Cesana-Arlotti, et al., 2012; Téglás & Bonatti, 2016). We next turn to examples that, from our perspective, illustrate how the more basic capacity of generating future possibilities from a present anchor can combine with other capacities to give rise to more complex capacities whose mastery shows a substantial degree of variation across phylogeny and ontogeny.

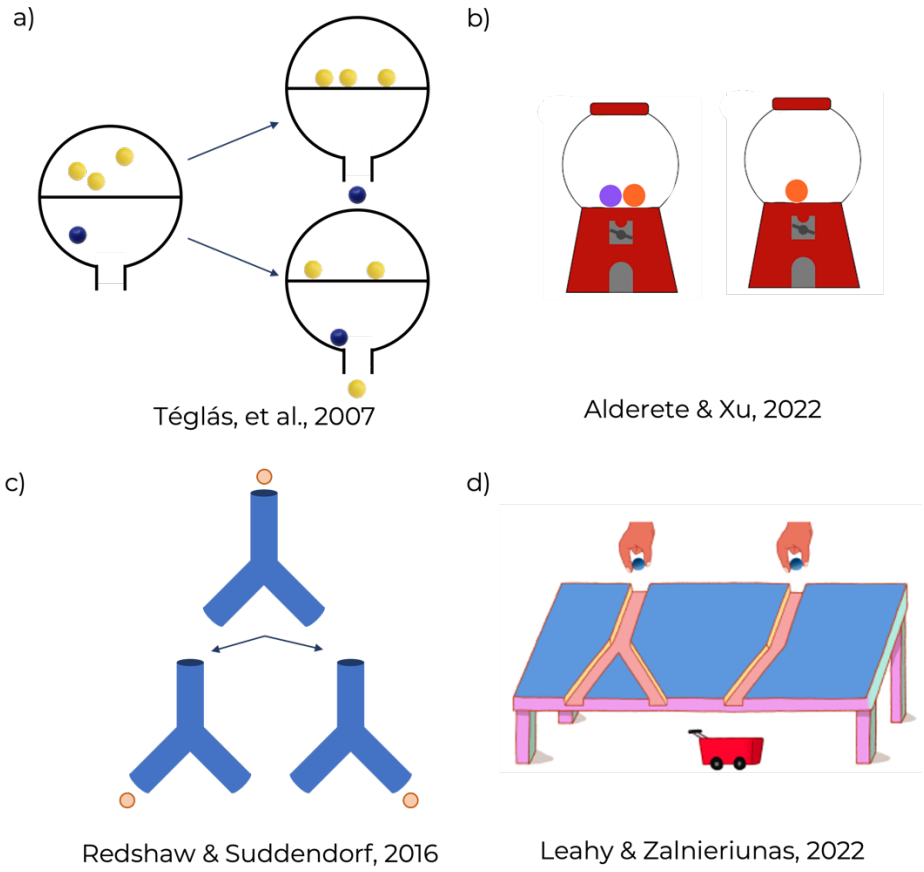


Figure 3. Illustrations of the experimental setup of four paradigms used to study the development of modal thought.

Making possible events actual

The same basic ability to generate future possibilities from a present anchor can be recruited in decision making tasks. Those tasks require future-oriented factual domain projection, but also the ability to compare possibilities according to their desirability, and, finally, the ability to use the results of the comparison as the basis for choosing the most desirable action. A recent study suggests this more complex suite of abilities is present by two and a half years of age, if not earlier. Alderete and Xu (2023), presented children with two gumball machines, one that had a single gumball in it, and another that had two differently colored gumballs, as shown in Figure 3b. What they found is that when seeking to get an orange gumball, for example, young children would choose the machine with a single orange gumball and avoid the machine with both a purple and an orange gumball. This suggests that the children were aware that choosing the machine with two gumballs might lead to a situation where they receive the undesirable purple ball. By at least the age of two and half, then, children do not only have the ability to represent possible future courses of events. They can also represent and evaluate the possible outcomes of different possible future actions and consistently select the preferred option.

Selecting actions conditional on possible future events

The results of the Alderete and Xu (2023) study stands in stark contrast with several studies that feature what looks like a closely related task—the 'forked-tube task' (see Fig. 3c)—which children struggle with until much later, and non-human primates do not pass at all (Beck, Robinson, Carroll, & Apperly, 2006; Leahy & Carey, 2020; Redshaw, & Suddendorf, 2020). In the forked-tube task, an object is dropped into a tube shaped like an upside-down Y, and subjects are incentivized to catch the object as it exits from the bottom of the forked tube (Redshaw & Suddendorf, 2016). Researchers have observed a remarkably robust pattern: non-human primates and young children before the age of 4 fail to systematically cover both

branches of the forked tube, instead alternating between covering one or the other of the two branching tubes (Redshaw & Suddendorf, 2016; Redshaw, Suddendorf, Neldner, Wilks, Tomaselli, Mushin, & Nielsen, 2019; Suddendorf, Crimston, & Redshaw, 2017).

At first glance this task appears quite similar to the gumball machine task. In both cases, children have a choice between an action that guarantees success and an action (or actions) that only has (or have) a 50% chance of success. In both cases, they have to represent possible future courses of events, evaluate them for desirability, and choose an action that leads to the most desirable outcome. But then given these similarities, why do children before the age of four robustly fail the forked-tube task while passing the gumball task with ease?

An important difference is that in the gumball task children are in control of which of the two possible future courses of events will end up being actual: whether or not they can be sure to receive the preferred gumball predictably depends on which action they take, and they are free to choose their action. The forked-tube task, in contrast, seems more demanding. Children are not in control of the two possible future courses of events they are presented with. Success now requires that they come up with an action that is optimal regardless of which of the two possible future courses of events will become actual. That is, they have to think of an action (or set of actions) that would allow them to catch the object regardless of which trajectory it may take. Covering the right side or the left side of the tube only sometimes results in catching the object, but only covering both guarantees success.

If this is the right way of thinking about why the forked-tube task is difficult, we can conclude (contra Leahy & Carey, 2020) that it's not because young children have difficulty generating multiple incompatible future extensions of the falling object. We think they understand that the falling object may both go left or go right, but they fail to see that covering both sides

simultaneously guarantees success, regardless of which trajectory becomes actual. Thus, the pattern observed in the forked-tube task in no way challenges the results of Téglás et al. (2007). Both involve future-oriented factual domain projection; it's just that the forked-tube task demands additional abilities as well. In line with this assessment about why younger children fail the forked-tube task, very recent work suggests that if children are first exposed to the different actions that may be taken (including covering both sides of the tubes), even three-year-old children begin to pass this task at remarkably high rates (Turan-Küçük & Kibbe, 2023). Similarly, if the task is altered such that the actions are more ecologically valid for non-human primates, there is now evidence that even chimpanzees pass a version of the forked-tubes task (Engelmann, et al., 2023).

Decomposing the abilities required by the forked-tube task in the way we have makes clear predictions about which kinds of tasks should be difficult and which should not be. For example, if young children are simply watching the object be dropped into the forked tube, it predicts that they should not be surprised if the ball exits out of the right or left side of the fork, indicating that they are representing both possibilities. Further, if the difficulty does arise from having to select an action that succeeds regardless of which of the two uncontrollable future possibilities will turn out to be actual, then even relatively small modifications to the gumball task (Alderete and Xu, 2023) should make it equally difficult for children to pass. Suppose, for example, that the experimenter in the gumball task explains that they will turn the handle on both gumball machines at once and the child's job is to simply catch an orange gumball as it exits. Young children should now be expected to struggle with this task in much the same way that they struggle with the forked-tube task. Success once again would demand that children select an action (or set of actions) that will succeed regardless of which one of two uncontrollable future possibilities will become actual.

Helpfully, an experiment that is structurally identical to the proposed modification of the gumball task has recently been conducted by Leahy and Zalnieriunas (2022). In this task (see Fig. 3d), children are shown two slides, one that has a single exit, and one that is shaped like an upside-down Y and thus has two exits. Two round objects are placed at the top of the two slides and will be dropped by the experimenter at the same time. Children's task is to place a cart underneath one of the three exits to catch one of the falling objects. What Leahy and Zalnieriunas found is that children younger than 4 place the cart under the single-exit slide only about 50% of the time. That is, in contrast to the gumball task, which children much younger than 4 pass with ease, they fail the seemingly similar slides task. The essential difference is that passing the slides task, like passing the forked-tube task, requires figuring out that some particular action maximizes success regardless of which of multiple uncontrollable future courses of events will be actualized.

To sum up, rather than thinking that children's failure on the forked-tubes task reveals a broad inability for modal thought, as argued by Leahy and Carey (2020) and Redshaw and Suddendorf (2020), we think it would be a mistake to draw conclusions from the results of this task in isolation. This task must be evaluated alongside other tasks that involve modal cognition, such as the Téglás et al. task and Alderete and Xu's gumball task. The project then becomes to explain what the broader set of success and failures suggests about which component parts of modal cognition are or are not present at a particular age or in a particular species.

A subsequent joint in modal thought: Epistemic anchors

A major consequence of our assumption that modal thought in its various instantiations relies on factual domain projection is that it predicts a critical hurdle that needs to be overcome before children master certain more difficult modal reasoning tasks around age 6. In all of the examples we have discussed so far, subjects engaged in modal reasoning could choose anchors from their environment that provided the factual basis for the possibilities they were interested in: situations with empty boxes, puddles, bouncing balls, gumball machines, forked tubes, and so on. Our account of modal thought in terms of factual domain projection predicts that there are certain kinds of scenarios where factual domain projection fails to generate the right domain of possibilities unless subjects choose anchors that include facts inside themselves – their own epistemic states. We will first discuss an example of such a scenario and then move on to show that modal reasoning tasks that children do not master before around age six involve precisely those kinds of scenarios.

Imagine you are climbing up a hill and see a man standing near a grove of trees in the distance. You can't yet clearly discern his face. Judging from his overall appearance, you think the man might be your friend Matt, but it might also be his partner Raphael. Unbeknownst to you, it happens to be Raphael. When we try to analyze your modal thought in the way we did before, we run into a problem. What actual situation should we take to be the modal anchor? Since the goal of actuality-directed modal thought is to figure out answers to questions about the world as a whole from an accessible part of it, the modal anchor should be an actual accessible situation that answers the question who the man might be. The situation you are looking at certainly qualifies. The example of the damaged box we discussed earlier showed that, for modal reasoning from anchor situations in the outside world, it shouldn't matter whether we do or do not know the relevant facts about it. But now we have a problem. The situation you are looking

at is a situation that has Raphael in it, not Matt. This means that in every possible extension of that situation, the person standing in the grove of trees will, of course, remain Raphael. There would be no possibility in the projected modal domain where that man could turn into Matt. Our account would thus seem to predict that you were wrong when you thought that the man you saw might be Matt, just as your friend was wrong when she told you that you could ship your pot in the box she pointed at. There is a clear intuition, though, that you were not wrong—for all you knew, the man standing near a grove of trees *might* be Matt! We need to find some way out of this dilemma.

There is another type of modal thought whose anchor is not located in the outside world but inside us. This type of anchor is our own current epistemic state, that is, the totality of our memories and perceptual experiences. We carry within us powerful mechanisms to perceive and represent facts about the world we live in and store the knowledge we have gained. Our perceptual experiences and memories are an invaluable resource that can be relied on when trying to figure out possible ways the world may be. In the example of the man standing near a grove of trees, then, *your current epistemic state*, that is, all your memories and perceptual experiences as you are looking at the man in the distance, can be used as a modal anchor. Your actual memories and perceptual experiences are part of the actual world, too, just as puddles and thunderstorms are. Possibilities from such an *epistemic* anchor would thus still be projected via factual domain projection as before.[4] All situations in the modal domain would have someone whose current memories and perceptual experiences exactly match yours. At least some of those situations can be extended to situations where that person is looking at

[4] This way of thinking about epistemic modal alternatives comes from Lewis (1996). For Lewis, your epistemic alternatives at a given time are the sets of worlds where you have the exact same perceptual experiences and memories you actually have at that time.

Matt. That is, when your perception of the man and what you know more generally about your two friends is compatible with the figure you see being Matt, there will be some possibilities in the projected domain where someone whose epistemic state matches yours is looking at Matt, in addition to possibilities where the person is looking at Raphael.

Modal reasoning from one's own epistemic state is an ability that is quite different from reasoning from anchors in the outside world. Yet both types of reasoning are actuality-directed, and modal domains for both can be generated by factual domain projection. The difference is the choice of anchor. The ability to reason from an epistemic anchor requires metacognitive abilities.

What is significant about the example of the man standing in the grove of trees in the distance is that it demonstrates how epistemic anchors—one's own epistemic states—can be *required* for certain kinds of modal thought. That means that passing certain kinds of modal reasoning tasks will *require* the cognitive ability to represent and reason from epistemic anchors. This insight is especially critical given that an important step in cognitive development is the relatively late emergence of the ability to reason about one's own epistemic state (Beran, Perner, & Proust, 2012; Kloo, Rohwer, & Perner, 2017; Rohwer, Kloo, & Perner, 2012). The example of the man standing in the distance also gives us a critical clue about which kinds of modal reasoning tasks do or do not require epistemic anchors. Our approach thus makes novel predictions about which modal reasoning tasks can be passed before, and which ones can only be passed after, the emergence of the ability to reason about one's own epistemic state.

The contrast between modal reasoning tasks that do vs. those that do not require epistemic anchors is brought out impressively in two tasks from Robinson and colleagues (Robinson, Rowley, Beck, Carroll, & Apperly, 2006) that are remarkably similar to one another. Children

who were somewhat older than those targeted in the forked-tube task (4- to 6-year-olds in this case) were introduced to two different paper bags—one of which contained only black building blocks, and one of which contained both orange and green building blocks. They were also introduced to a cardboard wall with three differently colored doors—orange, black, and green—and it was explained that building blocks would be pushed through the door that corresponded to their color. The children's task was to make sure that any building block pushed through the doors were caught by placing a tray beneath the door (illustrated in Fig. 4).

On some trials, called 'unknowable' trials, children were told that the experimenter was going to pick a block out of the bag with both orange and green blocks, and that they needed to make sure that the building block was caught when it came through the door. On these trials, Robinson and colleagues found that four- to five-year-old children placed two trays beneath the orange and green doors, indicating that they succeeded in generating some future possibilities in which an orange block was drawn from the bag and other future possibilities in which a green block was drawn. Critically, the pattern of responses diverged sharply on other trials, called 'unknown' trials. The only difference on these trials was that the experimenter had already drawn the block from the bag and placed it behind the corresponding door, but the children had not seen which block was drawn or which door it was placed behind. Now, when prompted to make sure the block was caught, four- to five-year-old children only placed one tray beneath either the green or orange door, but not two trays.

Given the minimal difference between the conditions, what explains why children were passing one version of this task but not the other? That is, why would children be better at generating the relevant possibilities for events that are 'unknowable' rather than simply 'unknown'? The difference is captured once one realizes that the 'unknown' task does, but the 'unknowable' task does not, *require* an epistemic state as anchor. In the 'unknowable' trials, situations that are

relevant for determining the possibilities we are interested in are situations where the actual blocks are still in the bag, and it's a situation of this kind that would naturally be chosen as the anchor. With such an anchor, future possibilities in the projected modal domain can differ with respect to the location of the block that would be drawn. In some future extensions of the anchor, a green block will be drawn and pushed through the green door; in others, an orange block will be drawn and pushed through the orange door. By the age of 4, children should be able to generate these two types of future possibilities from a present anchor situation and select actions that succeed regardless of which possibility will be the actual one (see Figure 4), which is exactly what Robinson and colleagues find.

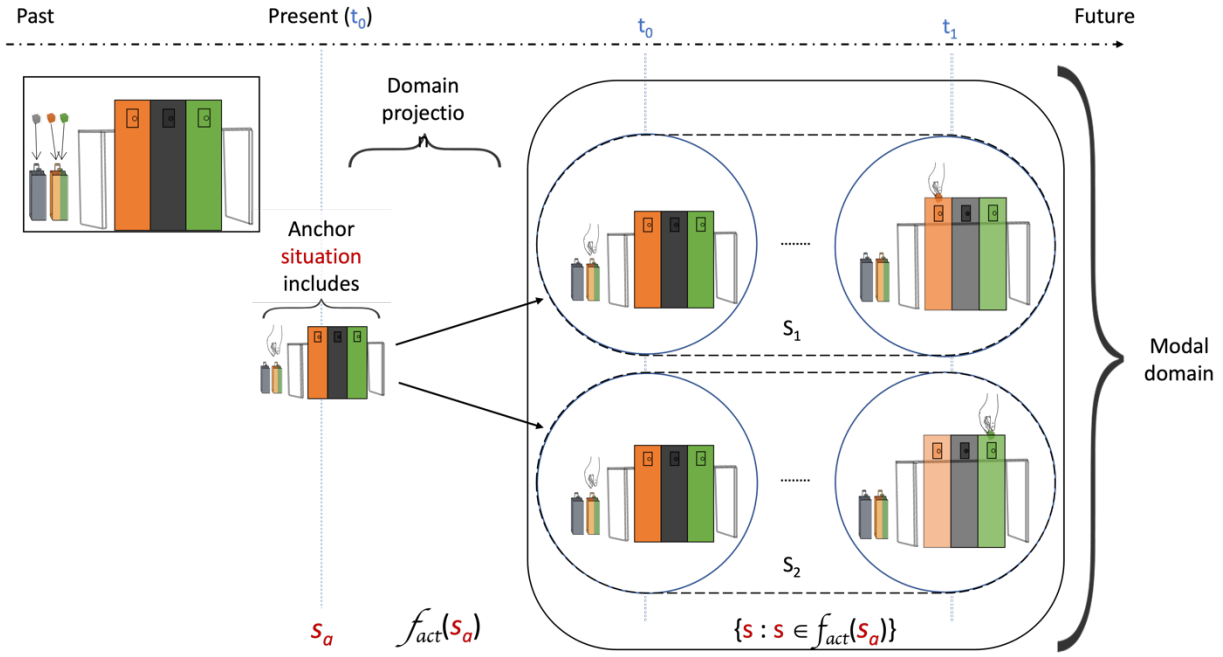


Figure 4. Schematic depiction of future-oriented factual domain projection used by children to pass the 'unknowable' trials in Robinson et al. (2006). A function f_{act} instantiates an inverse projection, taking as input the modal anchor s_a (an actual situation involving the experimental setup and the colored bags with blocks in them) and returning a set of situations, $\{s: s \in f_{act}(s_a)\}$, that have an exact match of s_a but also extend into the future. Among those situations are ones

in which an orange block is drawn and placed behind the orange door (s_1) and ones in which a green block is drawn and placed behind the green door (s_2).

However, in the 'unknown' trials, the strategy of choosing an anchor in the outside world will not work. It will not lead to a domain of possibilities that differ with respect to the color and the location of the block that has been drawn. On these trials, a block has already been drawn and has been placed behind the corresponding door. If we are looking for an anchor in the outside world, then these are the facts that are relevant for the possibilities we are considering. Yet if the modal anchor situation included those facts, all possibilities projected from it would contain the same block sitting in the same location. If 4- to 5-year-old children were assuming such an outside anchor, we could understand why they would only put down one tray. There would be no variation of blocks or their location across projected possibilities. If the actual anchor situation is a situation in which a green block is behind the green door, then in all future extensions of that situation, a green block will come through the green door. Alternatively, if the anchor situation is a situation in which an orange block is behind the orange door, then in all future extensions of that situation, an orange block will come through the orange door. Of course, the children don't know the color of the block that was drawn and behind which door it is sitting. So the best they can realistically do is pick a door at random. And this is what they seem to be doing.

This picture of what young children may be doing when failing the 'unknown' version of this task aligns nicely with what has been observed in a series of related studies involving finding a previously hidden object in one of two locations (Mody & Carey, 2016). As argued for by Leahy and Carey (2020), a promising interpretation of the data is that children are essentially selecting randomly between one of the two locations, but then taking themselves to be surprisingly certain that all relevant possibilities are ones in which that object is in that location (see also Leahy, Huemer, Steele, Alderete, & Carey, 2022). Similarly, this explanation also fits

neatly with the pattern of inferences found in Cesana-Arlotti and colleagues' work on infants reasoning about the identity of partially hidden objects (Cesana-Arlotti, Martín, Téglás, Vorobyova, Cetnarski, & Bonatti, 2018; Cesana-Arlotti, Kovács, & Téglás, 2020). Here too, the inferences children make would be expected if they are simply making a guess about the identity of the hidden object, and then updating their model of the actual world to be consistent with the additional information provided.

The question then is what it takes to pass the 'unknown' version of the task. Generating a modal domain that includes possibilities where the block that was drawn is behind the orange door *as well as* possibilities where the block that was drawn is behind the green door, can be achieved by using an epistemic anchor. If you use your own epistemic state— including what you see and remember and everything else you know about the situation—as the anchor, and you yourself do not know whether the block that was drawn and placed behind the corresponding door is green or orange, then you can generate present-oriented extensions of that anchor situation, and in some of those possibilities an orange block is behind the orange door, while in others, a green block is behind the green door (see Figure 5).

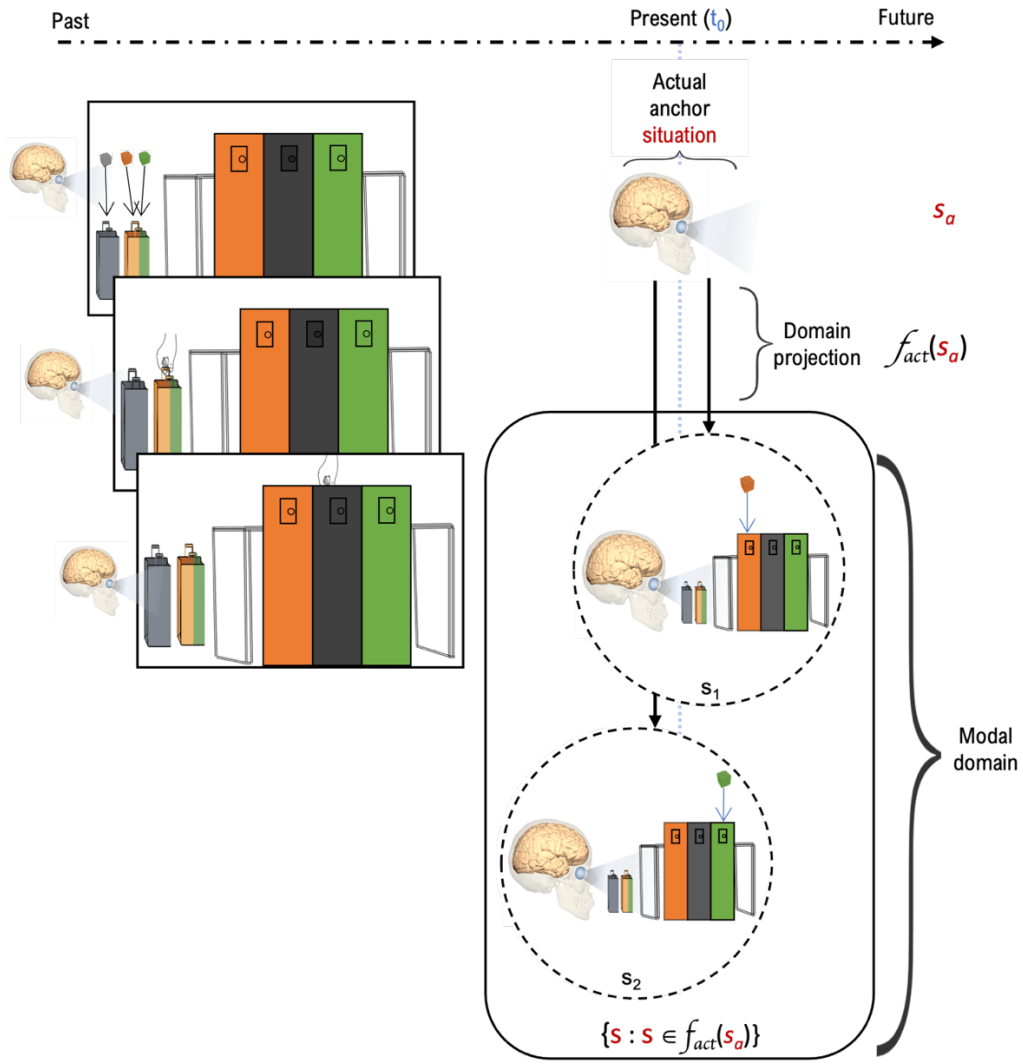


Figure 5. Schematic depiction of present-oriented factual domain projection required to pass the 'unknown' trials in Robinson et al. (2006). A function f_{act} instantiates an inverse projection, taking as input the modal anchor s_a (the situation in the actual world involving the participant and their actual epistemic state at the present time), and returning a set of situations, $\{s: s \in f_{act}(s_a)\}$, that have an exact match of s_a but also extend spatially. Among those situations are ones in which an orange block is behind the orange door (s_1) and ones in which a green block is behind the green door (s_2).

Given our explanation of the difference between the 'unknowable' and 'unknown' tasks, it should not be surprising that children pass the 'unknown' condition from Robinson and colleagues (2006) later in development. To be able to use your current epistemic state as the modal anchor requires that you can represent your own epistemic state independently from the outside world, and thus that you have the capacity for *metacognition about one's own epistemic state* (as a mere epistemic state). Accordingly, the relatively late development of the ability to solve tasks such as the 'unknown' condition tracks the relatively late emergence of the ability for this kind of metarepresentation (Beran, et al., 2012; Kloo, et al., 2017; Rohwer, et al., 2012).

Researchers directly studying this kind of metarepresentation have uncovered a similar distinction in the development of children's ability to reason about their own knowledge (Sodian & Wimmer, 1987; Rohwer, et al., 2012). In one study, Rohwer and colleagues hid a toy in an opaque box and asked 3- to 7-year-old children to assess whether they knew which toy was hidden in the box. When children were able to directly see the toy that went into the box, they correctly assessed the fact that they knew which toy was in the box. Similarly, when they had not seen any of the toys before one was hidden, they correctly assessed the fact that they did not know which toy was hidden (for similar findings, see, Pratt & Bryant, 1990; Tardif, Wellman, Fung, Liu, & Fang, 2005). However, when children were first introduced to a couple of toys (e.g., a train and a ball) before one of them was hidden in the box out of view, young children exhibited a remarkable difficulty in assessing their own knowledge: 70% of 3- to 5-year-old children claimed to know which toy was in the box (Rohwer, et al., 2012; see also, Sodian & Wimmer, 1987). Passing this task, just as passing the 'unknown' condition in the Robinson et al. task, requires the ability to reason about one's own epistemic state independently from the part of the external world in question. The external situation including the box is not one that is consistent with there being either a ball or a train in the box. It is either consistent with their being a ball and thus inconsistent with their being a train, or it is consistent with their being a

train and thus inconsistent with their being a ball. In contrast, children's epistemic states should be perfectly consistent with their being either a ball or a train in the box, and so their epistemic state does not settle what is in the box. Accordingly, if they have the ability to metarepresent and reason from their epistemic state independently from the external situation in question, they should be able to realize that they don't know what's in the box. If they can't, however, the best they can do is make a guess about which of those two possibilities is actual. Critically, children begin to pass both this task and the Robinson "unknown" task around the exact same age (~6 years), and they also fail both tasks in a similar way before the age of 6: in both, they act as if one of the two possibilities were actual.

Carving epistemic modal thought at its joints

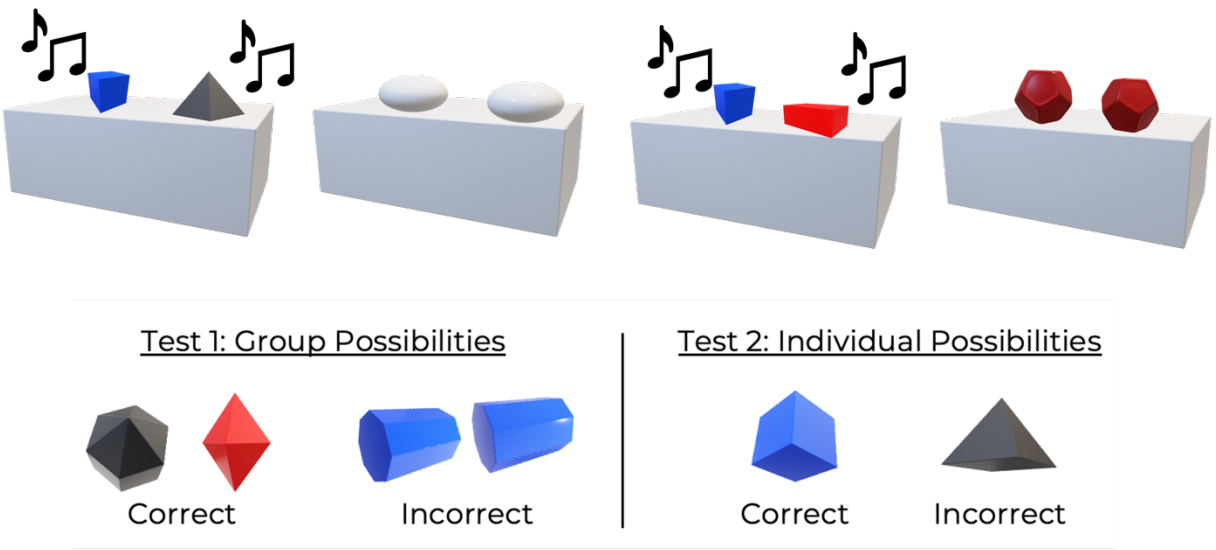
In the linguistic and philosophical literature on modality, cases like the one where you reason that a man standing in the distance on a hill might be some particular friend of yours are classified as instances of the same kind of modality as cases like the one where you come across a puddle and infer that it might have rained. Both types of inferences are standardly taken to be instances of epistemic reasoning, and both can be described using the modal 'might', which is commonly categorized as an 'epistemic' modal (see e.g., Papafragou, 1998; Cournane, 2020). On our account, there is an important difference between the two cases. The puddle-inference, unlike the distant-man-inference, does not *require* an epistemic modal anchor, hence does not *require* metacognition. Children who have not yet mastered metacognition about their own epistemic states should still be able to reason about possible past causes of present states or events, and thus should have no trouble reasoning about the possible causes

of a puddle.[5] By drawing a distinction between inferences that do and those that do not require epistemic anchors, we make novel predictions about which kinds of modal inferences children will struggle with before the age of six. Modal inferences that require epistemic anchors, hence metacognition, to pass should be mastered around the age of six, while modal inferences that do not require epistemic anchors, such as determining possible causes of an event, should be mastered much earlier.

A relatively large body of literature suggests that young children show surprising proficiency in entertaining multiple hypotheses concerning possible causes of some outcome (Gopnik, Sobel, Schulz, & Glymour, 2001; Gopnik & Wellman, 2012; Gweon & Schulz, 2011). One set of experiments by Goddu and colleagues (Goddu, Sullivan, & Walker, 2021) is particularly well suited to illustrate that causal reasoning does not necessarily require epistemic anchors, even though we might very well use so-called 'epistemic' modals to describe it. Children, between 18 and 30 months old, were given a causal reasoning task in which they had to determine what made a music box play music but were given insufficient evidence to identify a unique cause (see Fig. 6 top row). Specifically, children were shown evidence that was consistent both with (i) the box playing music whenever a mismatched pair of blocks was placed on top of it and (ii) the box playing music whenever one specific kind of block was placed on it (a blue cube in the illustration in Fig. 3e top row). All the evidence children received allowed for both of those two possible causes. After seeing the evidence, children were given two test trials that involved them selecting between blocks that could be placed on top of the music box to make it play music (Fig. 3e bottom row). On one test trial (Test 1: group possibilities in Fig. 6, bottom row),

[5] Of course, modal conclusions from puddles are ultimately risky when we don't know all the relevant properties of the puddle. For example, the presence of particular amounts of particular kinds of ions would exclude yesterday's rainstorm as having caused a particular puddle. In that case, an assertion that there might have been a rainstorm–given that puddle–would be false. There may be no serious consequences in this case, unlike in the case of the box with the damaged bottom.

children were presented with two *pairs* of blocks they had never seen before. One of the pairs had matching blocks and the other had mismatching blocks. On a separate test trial (Test 2: Individual possibilities in Fig. 6, bottom row), the same children were asked to choose between two individual blocks which they had never seen previously activate the musical box on their own. One block had been part of both mismatched pairs that had activated the box, while the other had only been part of one of the mismatched pairs. The order of these trials was randomized and children did not receive feedback on whether their choice was correct. Goddu and colleagues observed that the children correctly selected the mismatching pair of blocks rather than the matching pair, and the same children also correctly selected the individual block that had been a part of both mismatching pairs rather than the block that had only been part of one of the pairs (Goddu, et al., 2021). In other words, Goddu and colleagues found that even toddlers seem to be able to generate from a single piece of evidence multiple distinct possibilities involving different causes activating the box.



Goddu, Sullivan, & Walker, 2021

Figure 6. Illustration of the experimental paradigm used in Goddu, et al., 2021. The top row depicts the evidence young children observed about the conditions under which the box played music; the bottom row depicts the two separate tests young children passed.

As in the puddle case, we might use a so-called 'epistemic' modal for describing their modal thoughts in English: What made the box play *might* have been a mismatched pairs of blocks, but it *might* also have been the blue cube. Yet, crucially, this kind of experimental task doesn't *require* epistemic anchors to be passed. The modal domain could be projected from an actual anchor situation where the music box is playing. Possible extensions of the anchor situation considered in this case would be past-oriented. In some, a mismatched pair of blocks of any form or color was placed on the box, in others it was a matched or mismatched pair containing the particular block the children had seen activate the box before (in the setup in Figure 6, that would be the blue cube). While the modal reasoning required to pass the Goddu et al. music box task is complex, no epistemic anchors are required. On our account, the early success on even this relatively complex modal reasoning task is not surprising given the particular component pieces required to solve the task: Children could use conclusions obtained via past-oriented modal reasoning from anchors in the outside world to make decisions requiring future-oriented modal reasoning.

A late-emerging form of modal thought: Counterfactual reasoning

Perhaps the most well-studied form of modal thought in developmental psychology is the ability to reason *counterfactually*—an ability typically argued to not be fully operational until after the age of 8 and perhaps even as late as 13 (Harris, German, & Mills, 1996; Kominsky, Gerstenberg, Pelz, Sheskin, Singmann, Schulz, & Keil, 2021; Nyhout, Henke, & Ganea, 2019; Rafetseder, Cristi-Vargas & Perner, 2010; Rafetseder, Schwitalla, & Perner, 2013). Consider

the task that is currently taken to be the litmus test for counterfactual reasoning. In this task, known as the 'Muddy Shoes' task, children are told a story in which two children, Susie and Max, have been playing outside and have gotten their shoes muddy. Susie and Max then both walk into the kitchen without taking their shoes off, and the kitchen floor becomes dirty. After hearing about what happened, children are asked whether the floor would have still been dirty if one of the children, e.g., Susie, had taken her shoes off. Using this kind of counterfactual reasoning task, researchers have found that children typically don't pass it until quite late in development; they say that the floor would not be dirty in that case (Rafetseder, et al., 2010; Rafetseder, et al., 2013).

The modal thought required by this task differs in a number of ways from the kinds of modal cognition we have considered thus far. First, there is a difference in the *time* of the modal anchor. In all the previous cases we discussed, the modal anchor was a contextually salient present part of the actual world. In the Muddy Shoes scenario, potential modal anchors are all in the past. We are looking at possible worlds that match the actual world with respect to some past situation where Susie and Max already had muddy shoes but had not yet walked into the kitchen (Arregui 2009). Second, in counterfactual reasoning, one needs to generate possible future extensions of the anchor situation where an assumption one knows to be actually *false* becomes true: Contrary to fact, Susie has to be assumed to not walk through the kitchen with muddy shoes. Thus, possible anchor situations cannot include the fact that Susie walked in the kitchen and other facts that are inconsistent with her not doing so. In the other forms of modal thought discussed thus far, the range of possibilities considered was always compatible with one's understanding of the actual world—they were possibilities that, for all one knew, might have been actual. Finally, the possibilities for counterfactual reasoning have to stay *close* to the actual world; they can't stray too much from actuality. In the Muddy Shoes scenario, the 'correct' judgment would require Susie to take off her shoes and not leave mud on the kitchen floor, but

everything else should stay the same. In particular, Max would leave his shoes on and thus the floor would still get dirty. The pattern of errors observed in the Muddy Shoes task is that before the age of 8, children tend to say the kitchen floor would *not* have been dirty if Susie had taken off her shoes. The question we face is, given the suite of coordinated abilities that are required to pass this task, why are children's responses differing from those of adults?

It is unlikely that the difficulty arises merely from representing past anchors. By this late in development, children have more than sufficient short-term memory to be able to recall the events in the presented narrative (Gathercole, 1999). Moreover, the studies typically include memory check questions to make sure that children correctly recall the past events that occurred, and those questions tend to be uniformly answered correctly (see, e.g., Rafetseder, et al., 2013).

Another possibility that has been pursued is that children may not be able to represent possibilities that are inconsistent with their own understanding of the situation (Riggs, Peterson, Robinson, Mitchell, 1998; Peterson & Riggs, 1999). While this is possible, children by the age of 5 do seem to be able to successfully generate and reason over counterfactual states of affairs in other tasks. They can, for example, pass the false belief task, which requires predicting how an agent will act based on a representation of the world that is inconsistent with their own understanding (Perner, Sprung, & Steinkogler, 2004; Phillips & Norby, 2021; Wellman, Cross, & Watson, 2001), so it seems unlikely that this fully explains the surprisingly late development of counterfactual reasoning either.

A third way of explaining the observed pattern of errors is that children may make different background assumptions about which facts hang together. For example, children might assume that Max and Susie were coordinating their actions. Maybe Max generally copies what Susie

does, or maybe they were just doing things *together*. If so, this should have a substantial impact on how one reasons counterfactually. If Max is a copycat, for example, then if Susi takes her shoes off, so does Max. In that case, then, staying close to the actual facts (including that Max is a copycat) means that the floor would not be dirty.

Kratzer (1981, 1989, 2012) discusses several cases where facts are 'lumped together' and thus stand and fall together when a counterfactual assumption is made. As in the Muddy Shoes case, this leads to apparent violations of the 'stay-close-to-the-facts' strategy: The more facts are lumped together, the more facts have to be discarded under a counterfactual assumption, and that means counterfactual conclusions may depart further and further from the way things actually are. Kratzer (1981) reports judgments about counterfactuals that are quite similar to those of the children in the Muddy Shoes task, but do not feel erroneous in any way. In one of her examples, she imagines that two friends—Hans and Babette—are spending the evening together. They go to a restaurant, 'Dutchman's Delight', sit down, order, eat, and talk. We are then asked to suppose counterfactually that Babette had gone to a different restaurant, 'Frenchman's Horror', instead. The question is where Hans would have gone in that case. Attaching importance to the fact that Hans and Babette are spending the evening together, a natural answer is that Hans, too, would have gone to Frenchman's Horror. Giving this answer, we no longer seem to hold on to the fact that Hans actually went to Dutchman's Delight.

In the light of Kratzer's example, a possible explanation for children's non-adult judgments in the Muddy Shoes task might be that they, unlike adults who were given the same scenario, were working under the background assumption that Susie and Max coordinate their actions, a possibility not explicitly mentioned in the story they were told. Recent work by Nyhout, and colleagues (2019) supports this suggestion. They found that children performed significantly better in the Muddy Shoes task when Susie and Max walked on the kitchen floor at different

times and for different reasons. Strikingly, children in this case began to say that the floor would still have been dirty if Max had taken off his shoes, providing evidence that the reason that children answered the way they did was that they were lumping Susie's and Max's actions together, or put more simply, they assumed that they were acting *together* (Nyhout, et al., 2019).

The assessment that, in counterfactual reasoning, the primary difference between children and adults may be that children make different background assumptions is also consistent with recent studies on children's counterfactual reasoning about the trajectories of physical objects (Kominsky et al., 2021). These studies found that when asked to engage in counterfactual reasoning, most 4-year-old children identified the correct situation in the past to begin reasoning from and did consider future possibilities that were inconsistent with what they knew had actually occurred. However, the possibilities they considered deviated, often very substantially, from what had actually occurred (Kominsky et al., 2021). This result is consistent with the assumption that the main difference between children and adults may relate to background assumptions that were not explicitly mentioned.

This account differs in an important way from accounts that argue that children before the age of 6 do not engage in counterfactual reasoning *per se* but rely on "basic conditional reasoning" to pass some counterfactual reasoning tasks (Rafetseder, & Perner, 2014; Rafetseder et al., 2010; Rafetseder, O'Brien, Leahy, & Perner, 2021). In contrast, to this proposal, we suggest that children might very well master all the logical components of genuine counterfactual reasoning but may make different background assumptions than adults. Counterfactual reasoning, like most other types of modal reasoning, relies on background assumptions that are rarely ever stated explicitly. Crucially, they are not part of the meaning of counterfactuals *per se*. We hope

future work will provide tests for showing how children and adults may more generally differ in the way their reasoning relies on background assumptions that are not explicitly mentioned.

Stepping back from the details of particular studies on counterfactual reasoning, the point we've sought to illustrate is how even quite late-emerging forms of modal thought, such as counterfactual reasoning, are still built around the basic capacity for factual domain projection, though they also require additional capacities. Mastering the counterfactual reasoning tasks we discussed requires subjects to locate a suitable past anchor and to consider future-oriented possibilities projected from it. The special challenge presented by counterfactual reasoning is that the possibilities considered are known to be non-actual, but they can't diverge too far from actuality. What that means seems to at least partly depend on what background assumptions about the facts of the actual world we are making. That's a difficult issue, not just for a child. Counterfactual reasoning is known to be a slippery and highly context-dependent matter (see, e.g., Kratzer, 1981). As before, by decomposing the complex ability to reason counterfactually into its component parts, we can begin to pinpoint more precisely where children's capacities differ from those of adults.

Relation to alternative proposals concerning the development of modal thought

With the basics of our account on the table, and examples of how it works in place, we want to situate our account in relation to other proposals concerning the development of modal cognition.

Leahy & Carey; Redshaw & Suddendorf

One recent debate has focused on two competing explanations of the success children begin to have on modal reasoning tasks around the age of 4 (Carey, Leahy, Redshaw, & Suddendorf

2020), with Leahy and Carey focusing on the acquisition of modal concepts (2020), and Redshaw and Suddendorf focusing on the role of mental time travel (2020). At a broad level, our account differs in similar ways from both of these proposals. A major difference between our account and that of Leahy and Carey (Leahy & Carey, 2020; Leahy, et al., 2022), is that they propose a single kind of conceptual development that underwrites the general capacity for representing possibilities. In contrast, our account decomposes this general capacity to represent possibilities into component parts and argues that, as a consequence, this complex ability develops in a more piecemeal way. One straightforward place to see the difference between these two approaches is to consider the relatively late success children have on tasks that require an epistemic anchor. Recall the minimally different conditions from Robinson et al. (2006), discussed above. In the 'unknowable' condition, a block has not yet been drawn, and by the age of 4, children consider the possibility of two different kinds of block being drawn. In the 'unknown' condition, the only difference is that a block has already been drawn and placed behind a door, yet the same children no longer consider multiple possibilities regarding the kind of block it might be. According to our account, children can succeed on the 'unknowable' condition, in which the object has not yet been drawn from the bag, by using a non-epistemic anchor that includes the entire experimental setup; however, passing the 'unknown' condition requires an epistemic anchor, and thus the ability for metarepresenting one's own epistemic state. Accordingly, our account predicts a crucial difference in these two kinds of modal thought that is linked to a stage in cognitive development where specific kinds of metacognitive abilities emerge. For Leahy and Carey, by contrast, failure to pass either one of those tests is diagnostic of the lack of the *general* ability to simultaneously represent two mutually incompatible possibilities (Leahy & Carey, 2020; Leahy, et al., 2022). More broadly, because their account proposes that a radical developmental change happens around the age of 4, they have to offer alternative ways of explaining successes on modal reasoning tasks before that age (e.g., Alderete & Xu, 2022; Cesana-Arlotti, et al., 2012; Goddu, et al., 2021), as

well as failures after that age (e.g., Harris, et al., 1996; Kominsky, et al., 2021; Rafetseder, et al., 2010; Robinson, et al., 2006). Thus, an important theoretical advantage of our decompositional account is that it makes it possible to formulate testable hypotheses about possible developmental sequences that go beyond locating just one major juncture in the development of modal thought. Our account differs in similar ways from that of Redshaw and Suddendorf (2020), who propose that modal thought develops from the general capacities for "mental time travel" and recursion, and like Leahy and Carey (2020), argue that children before the age of 4 (and non-human primates) simply lack the ability to represent possibilities. While their account does seek to explain the comparatively late development of counterfactual reasoning (Harris, et al., 1996) in terms of the interaction between mental time travel and recursion, they similarly do not have natural ways of distinguishing between epistemic and non-epistemic modal thought, or of accounting for the successes that children have on modal tasks before the age of 4.

Harris

An alternative recent proposal by Harris (2022a) has argued that children represent possibilities well before the age of four, emphasizing that they clearly engage in pretense and imagination by the age of two (Harris, Kavanaugh, Wellman, & Hickling, 1993; Harris, 2022b). This early-emerging form of reasoning and talking about non-actual events is argued to *not* be found in non-human primates (Harris, 2022a). Accordingly, this proposal, like the previously discussed ones, holds that the ability to represent possibilities is unique to humans; however, unlike the previous proposals, it locates the emergence of possibility reasoning in the human lifespan earlier (in the form of imagination or pretense), and suggests that this ability for pretense must continue to develop in some way to eventually allow for the passing of more complicated forms of possibility reasoning, such as the forked-tubes task.

While we agree with Harris (2022a) that the emergence of possibility reasoning occurs before the age of four and continues to develop, his proposal diverges from ours in an important way. Throughout, we have been at pains to point out that we are offering an account of *actuality-directed* modal thought—thought about the possible ways the *actual* world may be—and we have treated this as importantly separate from other kinds of possibility reasoning, including mere imagination. Harris conjoins two forms of possibility reasoning where we see a critical distinction. We want to briefly explain why we think these two abilities should not be thought of as continuous with one another. First, at a theoretical level, the two capacities function differently. Pretense typically involves a deliberative and effortful process of unfurling a *single* non-actual possibility, while the kind of actuality-directed modal thought we have been considering often does *not* require effortful or deliberative processes and typically involves simultaneously considering *multiple* possibilities. Second, one cannot be *wrong* about the possibilities one considers during pretense. Suppose you pretend that a banana is a phone and that you suddenly get a call on it from your second cousin asking for a favor involving making 200 tuna sandwiches. Despite the far-fetched nature of that possibility, there is no fact in the world that could somehow show you made a mistake in considering that particular possibility. The laws governing the actual world do not impinge.

One benefit of our decompositional approach is that it allows us to say more precisely how the two forms of reasoning about possibilities differ from each other and how they are similar. A crucial difference between pretense and actuality-directed modal thought concerns the modal anchor. For actuality-directed modal thought from non-epistemic anchors, the anchor is a part of the outside world, and this means that, by making wrong assumptions about the anchor, your modal claims could end up being wrong. In our opening example, it was facts about a damaged bottom of a particular cardboard box that together with general facts about the world determined that you were wrong in thinking you could ship a heavy pot in the box. In contrast, in cases of

pretense, the anchor situation does not have to be actual—it can be a merely imagined situation, and it could be located in a merely imaginary universe. No need to worry about laws of nature, or even consistency. You *cannot* be wrong about what you pretend: you are the one who determines the facts of the situation you are imagining, and you are the one who determines how that situation could develop. The two kinds of reasoning thus seem to function quite differently at an abstract, theoretical level.

In addition, the evidence that Harris highlights actually provides good empirical reasons for treating the two capacities separately. The ability to engage in imagination or pretense does seem to be unique to humans and seems to come online later in development than actuality-directed modal thought (Harris 2022b). Human children are able to represent and reason from merely imagined situations by the age of two, while non-human primates do not seem to be able to do so at all. In contrast, as long as no epistemic anchors are required, the capacity for actuality-directed modal thought does seem to be shared with non-human primates, and likely already emerges in some form in humans by the first year of life (see *e.g.,* Teglas, *et al.*, 2007). Thus, from a phylogenetic and ontogenetic perspective, there is also reason to think of these as different capacities.

Another benefit of our decompositional approach is that we expect there to be ways of thinking about possibilities that mix components of imaginary and actuality-directed modal thought. We might consider purely fictional continuations of an actual situation, for example, that make no attempt to stay close to what is possible in the actual world. Or we may consider realistic continuations of merely imagined situations that test hypothetical scenarios as part of a scientific investigation (Williamson 2016). This kind of thinking is what Byrne (2007) calls the ‘rational imagination’. From our perspective, there could very well be forms of counterfactual thinking that are not anchored in the actual past, as we have assumed for the Muddy Shoes example above

but may take off from imagined situations that are the results of imagining minimal changes in an actual situation (McHugh, 2023). For example, we may be wondering about the air quality of the room we are currently in if there were a hundred more people in it. To reason about such a case, we would probably not go back to a past time when the room was just beginning to fill, and then project forward to consider possible futures where a hundred more people enter the room. Instead, we would likely rely on merely imagined situations where the room we are currently in has another hundred people in it.[6] What we call 'counterfactual thought' is unlikely to be a single fixed complex of abilities. More likely, it is a term that, like modal thought more generally, subsumes different kinds of complex thoughts built from more basic components. The consequences for cognitive development and the cross-linguistic expression of counterfactual thought still need to be explored. The (de)composition of counterfactual thought is by and large uncharted terrain.

Normality and Modal Thought

The account we have offered thus far has left a major issue unresolved. Recall that factual domain projection returns a set of possibilities with a match of the anchor situation. This provides one kind of restriction on the kinds of possibilities in a modal domain—they all must have matches of the anchor situation. However, there will always be a wildly large number of far-fetched possibilities that include a match of the anchor situation. To return to our example of considering what one could pack in a particular box when moving, there will be a truly enormous number of things that could be packed in that box; among other things, these may include a pet

[6] This kind of strategy may also be used when we reason about what would have happened in cases where laws of nature are violated, and other cases where there isn't an obvious past branching point to go back to, as, for example, in David Lewis's example of what would happen if kangaroos had no tails (Lewis 1973).

cat, another slightly smaller but equally empty box, and, almost certainly, a single lentil. Thus, the problem is that without some further specification, most of the possibilities in the modal domain will be useless—the sorts of things one really shouldn't be reasoning about. For factual domain projection to be useful, the possibilities we consider need to be constrained to those that merit consideration (Phillips, et al., 2019). Fortunately, this problem is short-circuited by the pervasive role that *normality* plays in shaping the possibilities that we consider in a quite general fashion.

Default restrictions for modal domains

The default to considering only normal possibilities is quite general, and can be found, not only in modal thought per se, but whenever possibilities are invoked in reasoning—whether one is processing the meaning of a sentence or simply making assumptions about some unobserved part of the world. Kahneman (2011, p.77) provides a (completely unintended but beautiful) illustration of how normality constrains the possibilities we consider on the basis of verbal descriptions of scenarios. Kahneman mentions an item from Frederick's (2005) Cognitive Reflection Test (CRT) that goes as follows:

If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to take 100 widgets? 100 minutes OR 5 minutes?

In a footnote, Kahneman informs the reader that what he takes to be the 'correct' answer is '5 minutes.' However, to arrive at this apparently correct answer, we have to assume that the possibilities considered in this case are limited to those where each of the 100 machines produces 1 widget in 5 minutes, and where moreover all 100 machines start working at the exact same time. We must exclude, for example, possibilities in which machines work

sequentially, machines passing along a partially completed widget to the next machine. None of these assumptions logically follows from the verbal description we are given. They are expected to be presumed as normal.

This example is just one illustration of the pervasive role that normality has in shaping cognition, from memory, to expectations, to causal reasoning (Kahneman & Miller, 1986; Pettit and Knobe, 2009; Hitchcock & Knobe, 2009). In fact, it even seems to impact visual inferences about the likely shapes of partially occluded objects (de Wit & van Lier, 2002). Given our (de)compositional approach, what we should then expect is that the default normality-based constraints found throughout cognition combine with factual domain projection in predictable ways. The combinatorial result is that the modal domains returned by factual domain projection will only involve normal possibilities, by default. Perhaps then it is not surprising that default normality constraints have also been shown to shape actuality-directed modal thought *per se*. It is well-established, for example, that the possibilities we generate tend not only to conform to the laws of physics, but also involve events that are statistically likely and actions that are morally good, rational, and conform to conventional norms (Baillargeon, 1987; Kahneman & Miller, 1986; McCloy & Byrne, 2000; Phillips & Knobe, 2018).

The presence of normality constraints on modal thought has been found as early in human life as has been tested (Baillargeon, 1987; Brown & Woolley, 2004; Chernyak, Kushnir, Sullivan, & Wang, 2013; Kalish, 1998; Lane, Ronfard, Francioli, & Harris, 2016; Phillips & Bloom, 2022; Shtulman, 2009; Shtulman & Carey, 2007; Shtulman & Phillips, 2018; Spelke, 2001; Stahl & Feigenson, 2015; Téglás, et al., 2007). For example, even very young infants' expectations about the movement of physical objects assume that physical generalizations concerning object continuity and solidity will not be violated (Baillargeon, 1987; Spelke, 2001; Stahl & Feigenson, 2015). Moreover, the previously discussed work by Téglás and colleagues has shown that

infants' expectations about future events were guided by that event's probability. Infants were more surprised (measured by looking time) when an event with .25 probability occurred than when an event with .75 probability occurred (Téglás, et al., 2007). Moreover, when somewhat older children are asked explicitly about the possibility of different events happening, they judge that improbable events (e.g., being given all of the candy bars in a store for free) are impossible until well after 5 years of age. Remarkably, young children also explicitly judge that events in which agents violate prescriptive norms to be impossible, saying, for example, that it is impossible for someone to do something wrong, like taking a toy from another child. Importantly, these findings do not depend on children's interpretation of the word *possible*. They also judge that such events would require 'magic' to actually happen (Brown & Woolley, 2004; Phillips & Bloom, 2022; Shtulman & Carey, 2007; Shtulman & Phillips, 2018). And neither do those findings depend on something about English or western cultures, as similar findings have been reported across different languages and cultures (Chernyak, et al., 2013; Nissel & Woolley, 2022). And, as we discussed in the prior section in counterfactual reasoning, the possibilities considered under counterfactual assumptions are implicitly restricted to things evolving normally in these counterfactual possibilities (Kahneman & Miller, 1986; Rafetseder, et al., 2010).

Normality-based restrictions on the possibilities considered have also been shown to operate in adult human cognition as a default (Phillips & Bloom, 2022). Recent work has found that the events that participants generate in open-ended decision problems are highly constrained to those that are likely to occur, involve rational and moral actions, and are believed to be normal (Hecht & Phillips, 2022; Srinivasan, Acierno, & Phillips, 2022). Moreover, these same features extend to adult's *default* intuitions of what is 'possible'. In one study, Phillips and Cushman (2017) asked adult participants to make judgments of possibility of various kinds of events, including ones that involved irrational and immoral actions. Importantly, participants were either forced to respond under time pressure, or were asked to respond after reflection. When

participants were forced to make judgments of possibility quickly (and thus their judgments could not be adjusted from a 'default' understanding of the possibility of an event), they more tended to judge events involving immoral and irrational actions as impossible (Phillips & Cushman, 2017).

This broad family of findings fit nicely with our (de)compositional approach and offers a way of understanding why the possibilities returned by factual domain projection will, by default, only involve normal possibilities. However, an important consequence of this aspect of our account is that the default constraint of normality is not actually an internal feature of the capacity for actuality-directed modal thought *per se*. That is, the core components of actuality-directed modal thought (factual domain projection from an anchor situation) do not guarantee the normality of all situations with an exact match of the anchor; rather, they allow for any situation—even abnormal ones—to be included in the domain, as long as they include a match of the anchor situation. Accordingly, our account predicts that while the possibilities reasoned over will, by default, be constrained to normal ones (a generally useful heuristic), this tendency should not be able to withstand explicit challenge by facts of the actual world. We provide an empirical test of this separation between factual domain projection and normality-based default constraints next.

Investigating the defeasibility of normality constraints

Consider the following scenario:

> **Scenario 1**: A child was born two years ago. The child was born from its mother's first pregnancy and its mother died a year later before becoming pregnant again.

And now, given that context, consider the modal claim (1):

(1) The child must not have any siblings.

Most likely, (1) will strike you as true. As (1) makes a universal modal claim, this means that all of the possibilities you represented were ones in which the child has no siblings. But such a homogenous domain is not given by factual domain projection; this homogeneity is achieved instead as a result of implicit normality constraints. And these implicit assumptions are defeasible: Consider the possibility that the mother gave birth to twins, triplets, or even more offspring from a single pregnancy. Once raised, such possibilities cannot continue to be excluded from the modal domain (despite their obvious abnormality) and once these possibilities have been included (1) should no longer strike you as true.

To systematically demonstrate the relationship between defeasible normality constraints and modal domains, we conducted a study in which half of participants were presented with Scenario 1 and asked to evaluate (1). They were subsequently asked to consider the possibility of twins, triplets, etc. and then asked to reevaluate (1). We expected relatively high agreement with (1) initially and comparatively lower agreement with (1) after we challenged the default normality assumptions by raising the possibility of twins, triplets, etc.

Critically, to show that this manipulation worked specifically because it challenged defeasible normality constraints, we would also want to compare agreement ratings in this case to those in a case where the possibility of twins, triplets, etc. would not have been implicitly excluded from the domain in the first place. To do so, the other half of participants were instead given the following scenario:

Scenario 2: A dog was born two years ago. The dog was born from its mother's first pregnancy and its mother died a year later before becoming pregnant again.

After reading Scenario 2, participants were asked to rate their agreement with (2).

(2) The dog must not have any siblings.

As with Scenario 1, participants were subsequently asked to consider the possibility of twins, triplets, etc. and then asked to evaluate (2) again. In this case, we expected relatively low agreement with (2) initially, suggesting that the domain may already include possibilities involving multiple offspring from a single pregnancy. Moreover, we also predicted that explicitly raising such possibilities should therefore also have less of an impact on participants' reevaluation of (2). To allow us to further investigate this proposed mechanism, we additionally asked both groups of participants whether they had considered the possibility of twins, triplets, etc. before we raised it.

Results. Participants' agreement ratings with the two modal claims showed the expected pattern overall (see Fig. 7a). Statistically this pattern can be captured by the significant interaction effect between whether the scenario concerned a child or a dog and the impact of raising the possibility of twins, triplets, etc. This interaction was highly significant using a comparison of linear mixed-effects models (χ^2 (1)= 13.967, p< 0.001). Specifically, in the case of the child, we found that participants largely agreed with (1) before the challenge (M= 65.66; SD= 35.74), but their agreement decreased markedly after the possibility of twins or triplets was raised (M= 33.59; SD= 29.92), t(250.26) = 7.844, p< 0.001, d= 0.973. By contrast, in the case of the dog, participants did not strongly agree with (2) even before the challenge (M= 37.86; SD= 41.85), and raising the possibility of multiple offspring from a single birth had a comparatively

small effect on their reevaluation of the modal claim (M= 26.67; SD= 36.68), $t(246) = 2.24$, p= 0.026, d= 0.284.

We next considered whether these patterns could be explained by whether or not participants considered the option of multiple offspring from a single birth before we raised it. First, we found that when the scenario concerned a child, the majority of participants (72%) reported that they did *not* consider the possibility before we raised it. In contrast, when the scenario concerned a dog, the majority of participants (66%) reported that they *did* consider the possibility before we raised it. This difference was highly significant using a chi-square test ($\chi^2 = 60.65$, $p< 0.001$). Second, combining the data from both conditions, we found that participants who had not thought of the possibility before we raised it overwhelmingly agreed with the modal claim, while those who had already considered this possibility, overwhelmingly disagreed with it. Accordingly, our explicit raising of this possibility only affected the agreement ratings of participants who had not already thought of the possibility (see Fig. 7b). Third, we conducted a mediation analysis that revealed that the differences in initial agreement with (1) vs. (2) were explained by differences in whether participants had considered the possibility of multiple offspring from a single pregnancy (95% CI of proportion mediated [0.43, 1.01], p< 0.001).[7]

A **B**

[7] These studies were approved by Dartmouth College's Committee for the Protection of Human Subjects (STUDY00032209). Additional experimental details, including stimuli, materials, data, code, sample sizes, participant exclusions, and a longer explication of the results can be found in the supplement to our paper: https://doi.org/10.7910/DVN/KUWNYK. This study was not preregistered.

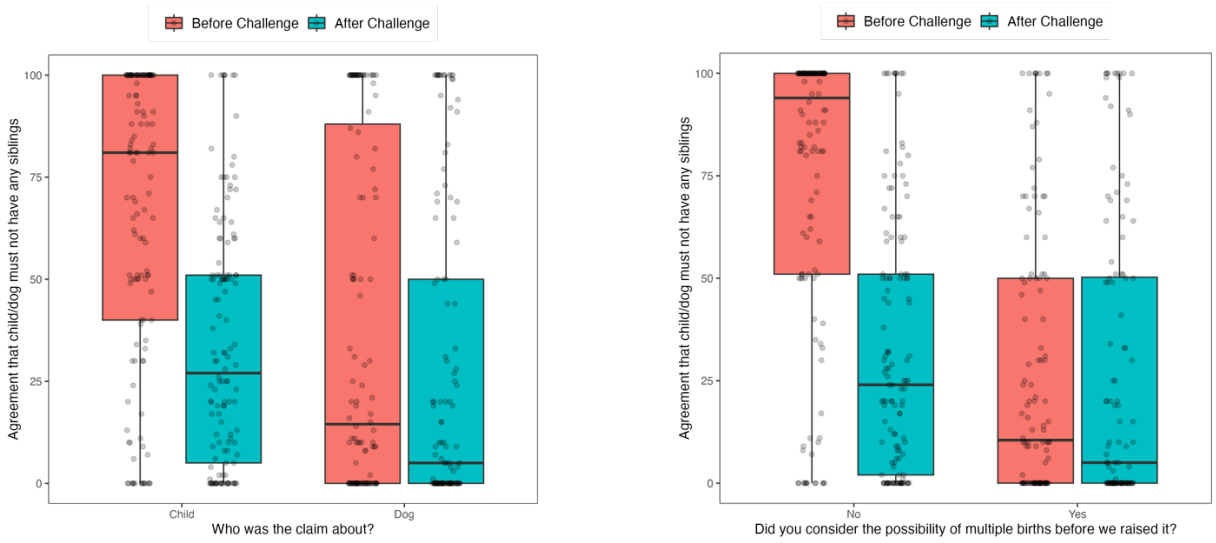


Figure 7. Boxplots of participants' agreement ratings with the modal claim, where small grey dots represent individual participants' agreement ratings. Red boxes depict agreement before we raised the possibility of twins, triplets, etc.; blue boxes depict agreement after we raised these possibilities. **A** The bars grouped on the left depict agreement ratings when the scenario concerned a human child; the bars on the right depict agreement ratings when the scenario instead involved a dog. **B** The bars grouped on the left depict agreement ratings of participant who indicated that they had not previously considered the possibility of twins, triplets, etc.; the bars on the right depict agreement ratings of participants who said they had considered such possibilities.

In short, what we find in this simple experiment conforms to the general contours of our account. On our approach, default normality constraints typically play a critical role in which possibilities are considered, but such default constraints are not special to modal thought. Moreover, these default constraints often do not stand up to challenge, suggesting that their influence does not come from the mechanism of factual domain projection itself.

(De)Composing modal meanings in natural languages

Introduction: Making the connection with language

In the first part of the paper we proposed a particular decomposition of actuality-directed modal thought centered around the basic ability of factual domain projection. We documented the explanatory value of the proposed decomposition by showing that it identified major milestones in the development of modal thought. If the core cognitive mechanism of actuality-directed modal thought is factual domain projection, we would expect that mechanism to be visible not only in developmental paths and cross-species variation, but also in the way modal thought is expressed within and across languages and in the way modal vocabulary is acquired. More specifically, we would expect languages to have systematic ways of representing the core ability of domain projection from an anchor, as well as the other abilities it combines with to produce increasingly complex forms of modal thought. Here are the component abilities of modal reasoning that we identified by looking at cognitive development:

> **Anchor choice:** The ability to choose an anchor containing the relevant facts for the possibilities considered.
>
> **Time of the anchor**: The ability to categorize anchors as present, past, or future.
>
> **Temporal orientation of possible extensions of the anchor**: The ability to represent possible extensions of the anchor situation as stretching into the future or into the past, or as staying within the present.
>
> **Epistemic anchors**: The ability to take one's own epistemic state as the anchor.

Counterfactuality: The ability to consider possible extensions of an anchor situation that one knows to be contrary to fact.

Restrictions for modal domains: the ability to navigate defeasible normality restrictions for domains of possibilities.

In what follows, we will briefly illustrate how those six components of modal reasoning are reflected in the combinatorial semantic systems of natural languages. Within the limits of this article, examples will mostly be drawn from English, but our discussion will be informed by what is more generally known about how, cross-linguistically, modal meanings are lexically expressed and how they interact with the contributions of other grammatical devices and building blocks, in particular with argument structure, temporal and aspectual operators in their vicinity (Chen et al., 2017; Rullmann & Matthewson, 2018), evidential markers (Aikhenvald, 2004, 2018), and mood (see Portner, 2018 for an overview). What we hope will emerge from this short and preliminary overview is that the meaning components that natural languages use to compositionally construct expressions of modal thought from smaller pieces match up fairly well with the components of modal thought we identified on the basis of various non-linguistic tasks that have been used to establish milestones for cognitive development.

In addition to identifying components of actuality-directed modal thought, the first part of this paper also delivered a major prediction of our approach that linked the ability for epistemic reasoning to the ability for representing one's own epistemic state. In this connection, we already pointed out that modals that are customarily classified as 'epistemic' in the linguistic and philosophical literature, do not necessarily require epistemic anchors. That is, their mastery does not necessarily require the capacity for metacognition. In this second part of the paper, we will come back to this fact and show that once we recognize that not all so-called 'epistemic

modals' require the capacity for metacognition, we will be able to settle an important and so far unsettled disagreement in the literature on the acquisition of so-called 'epistemic' modals and its relation to the emergence of metacognitive abilities.

Anchor choice: The arguments of modal words constrain the choice of anchors

An anchor-based semantics for modal expressions was first proposed by Hacquard (2006, 2010), and was developed further by Kratzer (2013). Kratzer hypothesized that modal anchors are generally provided by the arguments of modal words. The initial, still unrestricted, modal domain is then projected via factual domain projection from the anchor. Different kinds of modal words—*e.g.,* modal adjectives vs. modal auxiliaries—take different kinds of arguments, hence select different kinds of anchors, and thus project different kinds of modal domains. This variation can be used to get a glimpse of the role of anchors in modal language. By way of illustration, consider the following scenario from Lewis (1997:14; also Kratzer, 2013).

"A sorcerer takes a liking to a fragile glass, one that is a perfect intrinsic duplicate of all the other fragile glasses off the same production line. He does nothing at all to change the dispositional character of his glass. He only watches and waits, resolved that if ever his glass is struck, then, quick as a flash, he will cast a spell that changes the glass, renders it no longer fragile, and thereby aborts the process of breaking."

Against the background of Lewis's scenario, look at the following two sentences.

(1) The glass is fragile.

(2) The glass could/might break.

(1) and (2) are close in meaning. A glass that is fragile could break easily, hence could break. Yet we judge (1) as clearly true on Lewis's story but are more hesitant about (2). Modal adjectives such as 'fragile' have an individual argument that is realized as its subject. If modal anchors are generally provided by a modal expression's arguments, the subject of 'fragile' should provide the anchor. Thus, in (1), the modal anchor should be just the glass. We are considering possibilities that have matches of the glass at the current time, but the surroundings of the glass may differ in whichever way: the glass may be on a shelf or packed safely away, for example. Crucially, in the projected possibilities, the glass may or may not be protected by the sorcerer, and if it isn't, it may break at some future time (see Fig. 8). Thus, (1) seems true as there are future possibilities in the domain in which the glass breaks.

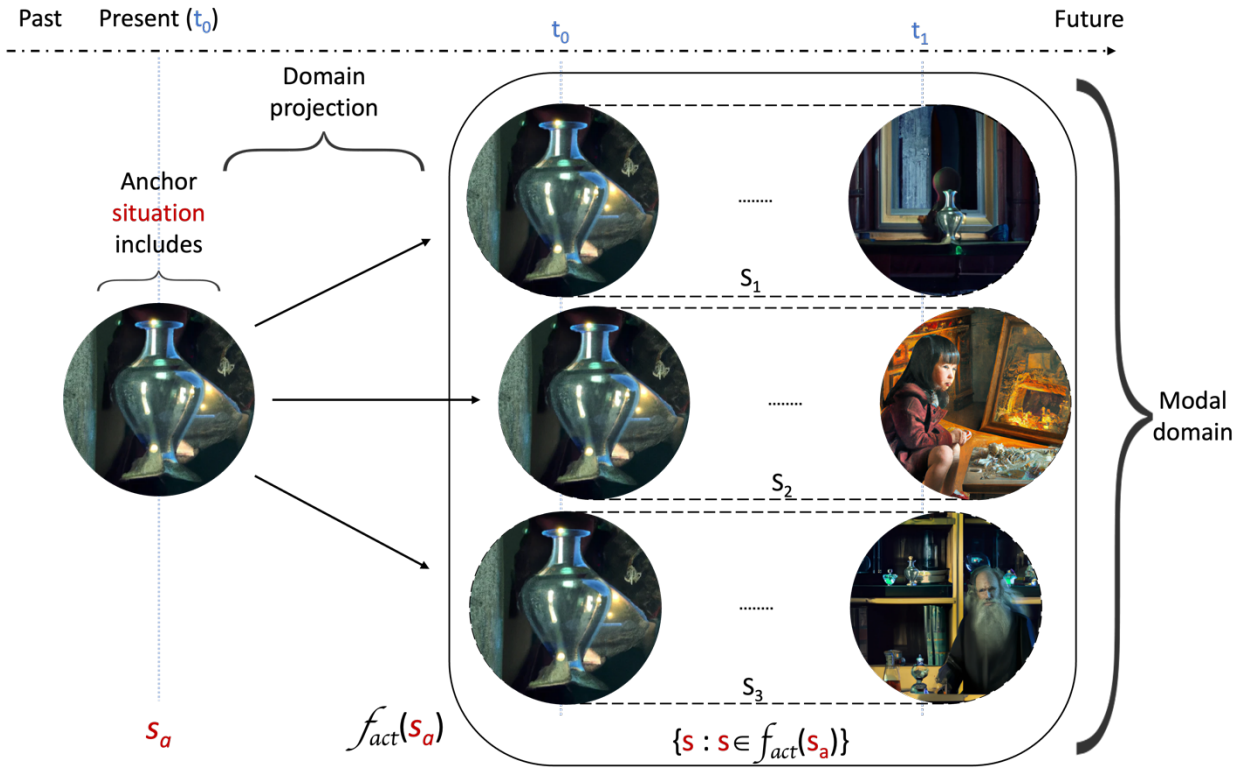


Fig. 8. Depiction of factual domain projection from an anchor that involves only the glass at the current time. Domain projection in this case returns a set of possibilites with exact matches of the glass at t_0. Among these may be situations in which it is not protected by a sorcerer but instead played with by a child and broken (s_2), as well as ones in which it is protected by the sorcerer (s_3). In situations that do not include the sorcerer, the glass may go on to break at a subsequent time, as in s_2.

Modal auxiliaries like 'could' are sentential operators. As sentential operators, they do not have individual arguments at all. Their only argument is a situation argument, which provides the evaluation situation for the modal statement as a whole. Contrary to first appearance, then, the grammatical subject in (2) is not an argument of 'could'. To see that, consider (3):

(3) Three glasses could fit on this shelf.

(3) is ambiguous. As expected with a sentential operator, there is scope interaction between 'could' and the quantifier phrase 'three glasses'. (3) might be understood as saying that there are three actual glasses that (individually or collectively) could fit on this shelf (wide scope of 'three glasses'). Or it might say that the shelf can accommodate three glasses—not any particular ones (narrow scope of 'three glasses'). The two readings can be represented by the logical forms 4(a) and (b), where 'could' is invariably a sentential operator. Being a sentential operator, 'could' may or may not take scope over the subject 'three glasses' (Bhatt 1999, Hackl 1998, Wurmbrand 1999).

(4) a. Three glasses λx (could (x fit on this shelf)).

'Three glasses have the property of being an x such that it is possible that x fits on this shelf'.

b. Could (three glasses fit on this shelf).
'It is possible that three glasses fit on this shelf'.

Since the auxiliary 'could' has no individual argument, the only plausible anchor is its evaluation situation. Situations, like locations more generally, but unlike individuals, are individuated poorly. The place where you live could be your house, your street, your neighborhood, town, county, state, nation, or planet. Likewise, possible anchor situations for 'could' may be smaller situations containing just the glass, or larger situations containing the glass together with the sorcerer. If, in the absence of any lexical constraints, modal anchors contain all the relevant facts for the possibilities considered, we expect to see variation in the choice of anchor depending on the kind of possibilities we are wondering about. To illustrate, if we are wondering about the potential of the glass all by itself, abstracting away from the particular situation it is currently in, we may argue that, of course, it could break: this is why it is being protected. But if we are asking ourselves whether the glass is vulnerable in the situation it is in, we may feel confident that it couldn't possibly break because it is protected at all times by a sorcerer. In the first case, we would choose a modal anchor that might just contain the glass. In the second case, we would include the sorcerer protecting the glass in the anchor situation. There are no grammatical pressures favoring or militating against either possibility. In this case, grammar doesn't tell us how to pick the situations that contain the facts that are relevant for the possibilities we are considering.

Returning to (2), the upshot is that the modal anchor for 'could' is more flexible than that for 'fragile'. Depending on the possibilities under discussion, the anchor for 'could' may

be a situation containing only the actual glass at the current time, or it may be a larger situation containing the current stage of the glass together with the sorcerer protecting it. If the anchor situation for 'could' includes the sorcerer, every projected possibility will have a match of that situation, and the sorcerer will thus be protecting the glass in each of those possibilities. So, if things go on to develop normally—the sorcerer doesn't quit or lose his powers—the sorcerer will protect the glass forever after, and the glass won't ever break (Fig 9).

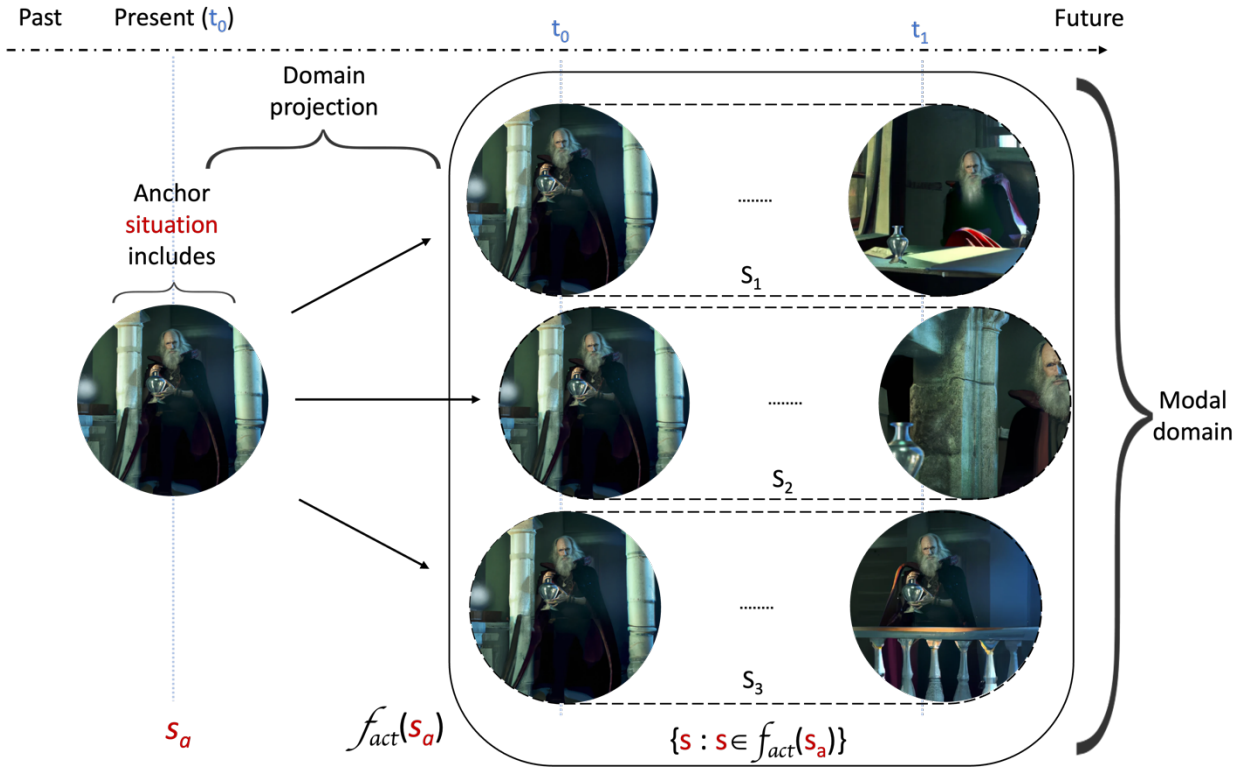


Fig. 9. Depiction of factual domain projection from an anchor that includes both the glass and the sorcerer at the current time. Domain projection in this case will return a set of situations with exact matches of the glass and sorcerer at t_0. In all of these situations, the sorcerer protects the glass, and thus, if things go on to develop normaly, the sorcerer will prevent the glass from breaking in all possibilities in the domain.

If the sorcerer isn't included in the modal anchor for 'could' in (2), modal domain projection proceeds as for (1), and there will again be extensions of the anchor situation where no sorcerer is protecting the glass and it breaks, as illustrated in Figure 8.

Casting the difference between modal adjectives and modal auxiliaries in this way, we can explain the hesitancy felt when evaluating (2): we might feel pulled by the thought that the glass couldn't break because it is protected by the sorcerer, or we might feel pulled toward the thought that the sorcerer is protecting the glass because it is fragile and thus could break. Due to this indeterminacy, we would expect truth-value judgments for (2) to be more variable than those for (1), and they should vary according to whether or not the sorcerer is taken to be part of the anchor situation. In contrast to (2), the modal anchor for (1) is entirely determined by grammar, and we should thus expect relatively more agreement in truth-value judgments.

Experimentally demonstrating the choice of modal anchors

To illustrate and experimentally test our predictions for (1) versus (2), we gave participants Lewis's original story and then asked them to rate their agreement with either (1) or (2).[8] Subsequently, we asked them whether they were (a) only considering the glass or (b) considering the sorcerer and the glass when rating their agreement with the modal claim in the context of Lewis's scenario. This experiment allowed us to test five key predictions from the account we have offered.

Results. First, we found that participants overall more agreed with the claim that the glass was fragile (M = 84.11; SD = 28.57), than with the claim that the glass could break (M = 61.58; SD = 39.29), confirming our original intuition, $t(514.08) = 7.626$, $p < .001$, d = 0.633 (see Fig. 10a). Second, participants' agreement ratings with (2) exhibited more variance than their agreement with (1) as predicted by the indeterminacy of the modal anchor for *could,* $F(328,201) = 1.892$, $p < .001$. Third, participants were less likely to report having considered the sorcerer when evaluating the 'fragile' claim (31.5%) than when evaluating the 'could' claim (64.5%), $\chi^2(1) = 52.91$, $p < .001$, which is predicted by the fact that the modal anchor for 'fragile' is given by the grammatical subject, while 'could' is more flexible in its choice of anchor and may or may not include the sorcerer (Fig 10b). Fourth, whether or not participants included the sorcerer in the modal anchor was predictive of whether they agreed with the modal claim, $F(1,524) = 156.4$, $p < .001$, $\eta^2_p = 0.176$ (Fig 10c). And finally, a mediation analysis confirmed that the difference in

[8] In this series of studies, we also asked participants to rate their agreement with modal claims about whether the glass is *vulnerable* and whether it *might* break. Agreement with these claims also fit the predicted pattern and are reported in the supplement. We do not report them here for simplicity and brevity.

agreement with (1) vs. (2) can largely be explained by whether or not the sorcerer was included in the modal anchor (95% CI of proportion mediated [0.32, 0,66], p< 0.001).[9]

[9] Additional experimental details, including stimuli, materials, data, code, sample sizes, participant exclusions, and a longer explication of the results can be found in the supplement to our paper: https://doi.org/10.7910/DVN/KUWNYK. This study was not preregistered.

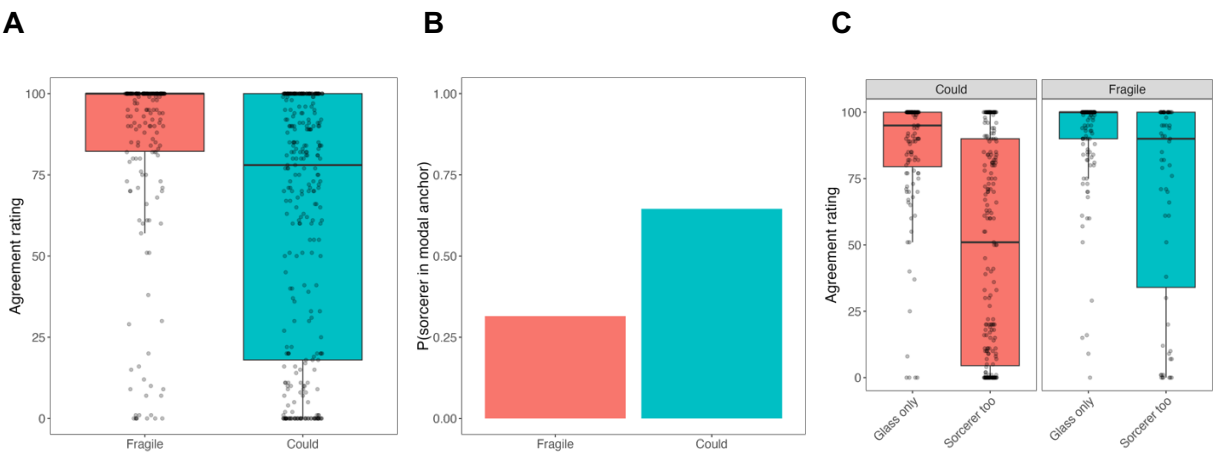


Fig. 10. **A** Boxplots of participants' agreement ratings with (1) (left box and points) and (2) (right box and points). Small grey dots depict individual participant responses, the colored boxes depict the middle two quartiles of responses, and the horizontal line depicts the median response. **B** depicts the probability that the sorcerer was included in the modal anchor for (1) (left bar) and (2) (right bar). **C** Boxplots of participants' agreement rating with (2) as a function of whether *only* the glass was included in the modal anchor (left boxes and points), or the sorcerer was included in the modal anchor along with the glass (right boxes and points), for both (1) right plot and (2) left plot.

Thus far in this section, what we've sought to demonstrate is how variation in truth-value judgments for different kinds of modal statements can be seen as arising from differences in modal anchors, where the choice of anchors is constrained by grammar. Modal domain projection from an actual anchor situation thus not only offers a way of accounting for variation in the development of modal reasoning, it also explains variation in truth-value judgments for different types of modal statements.

Anchor time and temporal orientation of possible extensions

In our review of the development of modal thought, we saw that the first attested instances of modal thought involve future-oriented domain projection from a present anchor. Not surprisingly, then, the earliest modals acquired by children are also present tense instances of future-oriented modals of the kind illustrated in (5) below, which are referred to as 'root' modals, as opposed to so-called 'epistemic' modals, in the linguistic literature (for overviews see Papafragou, 1998; Cournane, 2020).

(5) a. I **can** do this.

b. You **have to** help me.

In 5(a), the speaker says that there is a possible future extension of the current situation she is in where she does this action—whatever action it is she is referring to. In 5(b), the speaker conveys that in all possible—and acceptable—future extensions of the actual situation they are currently in, the addressee is helping her.

The modal anchors for the 'root' modals in 5(a) and (b) are located in the present. Natural languages also provide tools for talking about past modal anchors. For verbal modals with fully regular verbal paradigms the temporal location of the modal anchor is indicated by tense marking on the modal (Chen et al., 2017; Rullmann & Matthewson, 2018). (6) illustrates an instance of a future-oriented modal with a past anchor.

(6) I wasn't sure whether I **could** take on this responsibility.

In (6), the modal ‘could’ appears in its past tense form. The speaker is talking about a particular situation that occurred in the past. She is wondering whether there was a future extension of that situation where she takes on some particular responsibility. English thus uses tense morphology as a channel for specifying the time of the anchor situation.

That tense marking on modals indicates the time of the modal anchor, and not the time of the event or state described, is shown vividly by an example of Yalcin (2016). Yalcin imagines a man, Jones, who was buried during an earthquake, but has just emerged from the rubble as the only survivor. Now he is talking to the media. Imagining such a scenario, we see a stark contrast between 7(a) and (b).

(7) a. I **should** be dead right now.

b. I **must** be dead right now.

For Jones to say 7(a) in this situation is appropriate, 7(b) would be absurd. If ‘should’ is the past tense of the future-oriented modal ‘shall’, but ‘must’ is present tense and present-oriented in this case, we have an explanation for the contrast between 7(a) versus 7(b). With 7(a), Jones is saying that, given the past situation he was in – buried under the rubble – he would have been dead by now in all normal future extensions of that situation. In contrast, 7(b) would have conveyed that all present extensions of the situation he is currently in are situations where he is dead, which is absurd.

The modals in 5(a), (b), (6), and 7(a) are all future-oriented. So are all adjectival modals like ‘fragile’, ‘feasible’, ‘stealable’, and so on. A fragile glass is one that has a possible future instantiation that breaks, a feasible project has a possible future instantiation that is realized, and an stealable bike has a possible future instantiation that gets stolen. Strikingly, there do not

seem to be adjectival modals that have a past orientation. For example, despite its potential usefulness, we do not have an adjective ‘kleptose’ in English that is similar to ‘stealable’, but which applies to objects if they have possible past instantiations in which they were stolen. A bike would be kleptose if it is possible that in the past someone stole it, as in, ‘I wouldn’t recommend buying that bike, it’s kleptose’. In natural languages, much like in cognitive development, future orientation seems to serve as a default.

Chen et al.’s (2017) and Rullmann & Matthewson’s (2018) cross-linguistic investigations showed that, when expressed overtly, the temporal orientation of modals is encoded, not by the modals themselves, but by aspectual morphology associated with the embedded verb. (8) illustrates.

(8) a. She might **be** wait**ing** for you.

b. She might **have** depart**ed**.

In 8(a), the temporal orientation of the possibilities considered is present, which is indicated by the progressive form of the embedded verb *‘*wait*’*. 8(a) talks about possibilities where she is presently waiting for you. In 8(b) the temporal orientation of the possibilities considered is past, which is indicated by perfect aspect marking (‘have’ plus past participle) of the embedded verb ‘depart’. While future orientation of modals is not marked in English, this is not universally so. For example, a modal’s future orientation must be marked overtly on the verb embedded under the modal in Gitksan, an endangered Tsimshianic language of northwestern British Columbia (Matthewson, 2013).

To summarize, natural languages may explicitly encode both the time of the anchor situation and the temporal orientation of possible extensions of the anchor situation. Those two meaning

components are carried by tense and aspect markers respectively. The two markers contribute to a combinatorics of meanings that allows us to talk about present-, past-, or future-oriented possibilities projected from present, past, or future modal anchors.

Epistemic anchors and the 'epistemic gap'

In our review of the cognitive development of modal reasoning, we argued that a major developmental achievement is acquiring the capacity to use one's own epistemic state as an anchor for projecting possibilities. A natural question is then whether this distinction maps onto the much-discussed linguistic distinction between so-called 'root' and 'epistemic' modals (Papafragou, 1998; Cournane, 2020). (9) and (10) illustrate.

(9) So-called 'root' modals

I **can** (**must**, **may**, **will**, **should**) leave the room.

(10) So-called 'epistemic' modals

She **could** (**must**, **might**, **will, should**) have left the room.

So-called 'root' modals are always future oriented and relate to abilities, potentials, intentions, or obligations, while so-called 'epistemic' modals need not be future-oriented, but can also be past- or present-oriented, and have traditionally been understood as relating to information states or knowledge. Children produce 'root' modals from around age 2, and so-called 'epistemic' modal verbs from around age 3 (for overviews see again Papafragou, 1998; Cournane, 2020). In the Bristol Language Development Study, for example, agent-oriented uses of 'can' and instances of 'will' used for communicating intentions are the earliest uses of modals by children (Wells, 1979). So-called 'epistemic' uses of modals like 'may*'* and 'might*'* appear at age 3;3 (Wells,

1985). This difference in the time of acquisition for ‘root’ versus ‘epistemic’ modals is called the ‘epistemic gap’ in the language acquisition literature.

Papafragou (1998) entertained the hypothesis that the ‘epistemic gap’ occurs because ‘epistemic’ modals require metacognition, that is, the capacity for representing epistemic states. This hypothesis does not fit with the results of Kloo, et al. (2017), or Rohwer, et al. (2012), establishing that a capacity for metarepresentation of one’s own epistemic states emerges quite late and is perhaps only fully realized by age 6. Papafragou’s hypothesis has also been challenged by Cournane (2021; see also van Dooren et al. 2022). Cournane presents corpus data from 17 English learning children showing that they are using contextually appropriate sentences with ‘epistemic’ adverbs like ‘maybe’ from before age two, when they are still only using root modal auxiliaries. For her, the later appearance of so-called ‘epistemic’ modal auxiliaries, as opposed to ‘root’ modal auxiliaries, has a linguistic explanation, related to syntactic complexity.

Our approach to modal semantics via factual domain projection has the potential to resolve the conflict between Papafragou (1998) and Cournane (2021), while also being consistent with the results of Kloo, et al. (2017) and Rohwer, et al. (2012). We saw that not all instances of so-called ‘epistemic’ modals require epistemic anchors, that is, the metacognitive ability to represent one’s own epistemic state. For example, among the early cases of ‘epistemic’ adverbs mentioned by Cournane (2021: 221;) are examples like 11(a) and (b).

(11) a. **Maybe** grandma made this.

b. **Maybe** that’s a fish.

A possible context for 11(a) may be a situation where we are looking at a hand-knit sweater. We can easily imagine possible past-oriented extensions of the present actual sweater situation where a whole variety of people might have knit that sweater. No metacognition is required. 11(b) might be about a sketchy drawing. Here, we can easily imagine possible past-oriented extensions of the actual situation with the drawing in it, but where the drawing was intended to represent a sock or a potholder. Again, no metarepresentation of one's own epistemic state is required. Put in general terms, our factual domain projection approach predicts that metacognitive capacities are only required with those instances of so-called 'epistemic' modals where the choice of a non-epistemic anchor would yield modal domains that are not diverse enough, that is, domains of possibilities that all give the same answer to the question of our inquiry. The modal domains projected for 11(a) and (b) do not suffer from this problem. Even if grandma actually made that sweater, that's not a property that the sweater has in all possible past-oriented extensions of the anchor situation. It could have been made by somebody else. Likewise, even if this is actually a drawing of a fish, this is not a property that the drawing has in all possible past-oriented extensions of the anchor situation. It could have been created to represent something else.

On our account, Papafragou (1998) is right in assuming that there is a connection between mastering so-called 'epistemic' modals and the capacity for metacognition. Certain instances of so-called 'epistemic' modals do indeed require metacognition. Interestingly, the standard tasks in experimental studies on the acquisition of 'epistemic' modals are hidden-object tasks, and the kind of scenarios where the modals appear in those tasks are thus precisely the kind of scenarios that require epistemic anchors on our account. In Noveck et al. (1996), for example, children were introduced to two open boxes and a third, closed, box. They saw that one of the open boxes contained a bear and a parrot, and the other open box only had a parrot. They were told that the closed box had the same content as one of the open boxes, and then heard each of

two puppets articulate a modal statement. One of the statements was true, the other was false. The children's task was to say which of the two puppets was right. Among the contrasting statements that Noveck et al. tested were 12(a) and (b).

(12) a. There **might** be a bear in the box.

b. There **has to** be a bear in the box.

Given the information the children were presented with, 12(a) is true, but 12(b) is false. Yet even 5-year olds' choice between 12(a) and (b) was at chance. On our approach, this result is expected. To generate a modal domain where a bear is in the box in some, but not all possibilities, children would have had to use their own epistemic states as anchors. But the capacity for metacognition about one's own epistemic state only emerges around 6 years of age. If children instead use the actual box itself as the modal anchor, then they will either generate a domain in which all relevant possibilities are ones in which there is a bear in the box, or they will generate a domain in which none of the relevant possibilities are ones in which there is a bear in the box. In the first case, both sentences are true, and children should choose at chance if forced to choose between 12(a) and (b); In the second case, both sentences are false, and children should again choose at chance between 12(a) and (b). In short, our account predicts that they will not be able to distinguish between possibility and necessity statements in tasks that require an epistemic anchor.

While hidden object tasks allow us to detect children's inability to use epistemic anchors, there are many more situations where epistemic anchors are required: An approaching man might or might not be your friend Matt, a bird you are looking at might be a sparrow or a female house finch, and the plants you see on the shelf over there might or might not be fake. Do children use 'epistemic' modals in those situations? And if they do, how do they understand them? That last

question is critical, but we don't have the data to answer it. On our account, we would expect that children younger than 6 might very well produce modal statements like 'maybe the man is Matt', 'maybe the bird is a sparrow', or 'maybe the plants are fake', but, in line with Noveck et al. (1996), we wouldn't expect young children to be able to distinguish between possibility and necessity interpretations for those statements. In other words, we would expect children to understand that these claims are modal claims, but we would also expect them to agree with 'this bird might be a sparrow' to the same extent as with 'this bird has to be a sparrow'.

To conclude, from our perspective, the traditional, pretheoretical, category of 'epistemic' modal covers two uses that can – and should be – distinguished, depending on whether or not they *require* an epistemic anchor. The distinction is crucial for understanding a critical juncture in the development of modal thought, as we have seen. However, while the distinction is critical in cognitive development, the modal systems of languages like English do not mark it overtly. There do not seem to be any English modals that require epistemic anchors. English thus does not seem to have genuine epistemic modals.

A type of linguistic expressions whose mastery truly requires metacognitive capacities about one's own knowledge are so-called 'evidentials'. Evidentials are grammatical markers that represent the way evidence for the truth of an assertion was acquired. Evidentials can be found in many unrelated languages around the world. They represent the type of evidence a speaker has for the truth of a proposition (for overviews see Aikhenvald, 2004, 2018; Murray, 2021a, 2021b). In some languages, evidentiality is obligatorily marked in every sentence. Tariana, an endangered Arawak language spoken in Northwestern Brazil, for example, has obligatory verbal inflections that mark whether the evidence supporting an assertion was direct visual evidence, direct auditory evidence, indirect visual evidence, evidence via common knowledge inference, or hearsay (Aikhenvald, 2004). There are also languages that have evidential tenses, that is,

tenses that mark the time the evidence for an assertion was acquired. Korean and Mvskoke (Creek), an endangered Eastern Muskogean language spoken in Oklahoma, are examples (Lee, 2013; Johnson, 2023).

Mastery of evidentials requires the capacity to reflect on one's own epistemic state and on the source of relevant information it contains. It should thus not come as a surprise that evidentials are acquired late. Ozturk and Papafragou (2016), for example, found that the semantics and pragmatics of evidential morphology in Turkish are only acquired at around age 6 or 7. This fact aligns perfectly with the age when children begin to succeed on tasks that require reasoning about their own mental states.

We conclude that the grammars of natural languages have the means to overtly represent modal reasoning from one's own epistemic state, but, it is evidential markers, not multi-purpose modals like English 'might' or 'must', that unequivocally indicate such reasoning.

Counterfactuality

The expression of counterfactuality across languages is more complicated and more varied than the other modal constructions we have looked at (Arregui, 2021; von Fintel & Iatridou, 2023).[10] This would be expected if, as Arregui suspects, the term 'counterfactuality' doesn't stand for a

[10] Linguistic and psycholinguistic investigations of counterfactuals usually look at constructions where the counterfactual assumption is expressed by a subordinate sentence corresponding to an *if*-clause in English. But there are, of course, also counterfactuals like (i):

(i) Without a coat you would be freezing now.

These constructions may offer researchers in Cognitive Development a promising method for investigating counterfactual reasoning independently of a demanding syntactic construction.

unified phenomenon but is an umbrella term subsuming different kinds of complex modal thought. Still in many languages, we can detect overt realizations of some of the components of counterfactual thought we identified earlier (von Fintel & Iatridou, 2023). As an illustration from English, consider (13) against the background of the Muddy Shoes scenario discussed previously:

(13) If Susie had taken off her shoes, Max **would** have, too.

In the context of Muddy Shoes, we readily understand (13) as making a counterfactual assumption. After all, we were told that Susie and Max did *not* take their shoes off. Where does the counterfactual interpretation come from? It seems that English does not have dedicated morphology contributing a counterfactual interpretation.[11] (14) makes that clear.

(14) Copycat

When I came into the kitchen, I noticed that Susie had taken off her muddy shoes. I couldn't see Max, but knowing what a copycat he is, I knew that **if Susie had taken off her muddy shoes, Max would have, too**.

(14) provides a context for (13) where it is not interpreted as a counterfactual. Whether a sentence like (13) does or does not have a counterfactual interpretation, then, depends on the

[11] The only remnant of what might be taken to be counterfactual morphology in English is the form *were* in cases like 'if I were here, you would be, too.' Yet even this form doesn't necessarily signal that an assumption is false. It may also signal uncertainty about the truth of an assumption, as in: 'If I were chosen for the position, I would turn your company around.' This could be a promise made in a job interview by someone who still hopes to be chosen for the position. See von Fintel & Iatridou (2023) for more discussion.

context where it is used. (13) acquires a counterfactual interpretation only in contexts where it is presumed that Susie didn't take off her muddy shoes. In the context of Muddy Shoes, the *if*-clause (antecedent) of (13) is known to be false, in the context of Copycat, it is known to be true. The same conditional (13) can also be used in contexts where it's unknown whether its antecedent is true, as in (15).

(15) Uncertainty

Susie and Max were playing outside and I saw them enter the house with their muddy shoes. I feared that they might have entered the kitchen without taking their shoes off. Knowing Susie as the more reasonable kid, I thought there was a slight chance that she might have taken her shoes off before going into the kitchen. And, knowing what a copycat Max is, I was sure that **if Susie had taken off her muddy shoes, Max would have, too**.

While English has no distinctive marking of counterfactuality, two of the components of counterfactual thought we identified earlier are expressed overtly: the pastness of the modal anchor and the future orientation of the possibilities considered. The future-oriented modal 'will' appears in its past tense form 'would' in (13), and the pastness of the modal anchor is also reflected by the past tense appearing in the antecedent. Unlike English, German, for example, uses past subjunctive mood to mark the possibilities considered not only as anchored in the past, but also as uncertain or contrary-to-fact.

We mentioned earlier that there is a kind of counterfactual thought where we don't consider unrealized possible futures branching off from some point in the actual past of our world, but reason about the effects of instant counterfactual changes in an actual current situation. Those two types of counterfactual reasoning are not expressed differently in English, or, as far as we

know, in any other language. In English, for example, we use the past tense form 'would' of the future-oriented modal 'will' in both cases.

(16) If there were a hundred more people in the room right now, the air **would** be unhealthy.

That the same morphological devices can be used for the realization of both kinds of counterfactual thought is not surprising given our anchoring semantics for modals, which, in the case of counterfactuals, amounts to the account of Arregui (2009). On this account, past tense marking on the modal 'would', like past tense marking on the modal 'should', does not indicate pastness of the event or state being described, but, crucially, pastness of the modal anchor, which provides the relevant facts for the modal claim. For (16), for example, the modal anchor could be a situation that includes the actual state of the room in question during a time interval that includes the current time as its endpoint. We are imagining a change of the anchor situation at the current time and considering possible future extensions of the resulting changed situation(s). The modal anchor would still be past for this kind of reasoning (a situation beginning in the past and ending at the current time), and the possible extensions considered would still be future oriented (beginning at the current time and extending into the future from there). Past tense marking of the future-oriented modal 'will' is thus expected even in cases where we are imagining an instantaneous counterfactual change of a current situation. There is, of course, still a question why in sentences like (13) or (16), past tense does not only show up on the modal 'would', but also on the verb in the 'if'-clause. Various explanations, some purely syntactic, have been given for this fact.[12]

[12] See Arregui (2021) for a short discussion and references. To see that the past tense in the 'if'-clause of (13) or (16) isn't specific to counterfactual constructions and why it may be considered a syntactic phenomenon, consider (i), which is not a counterfactual, but has a syntactic structure that is very much like that of counterfactuals. In (i), the verb 'leave' in the subordinate clause appears in its past tense form even though the event being described is in the future.

Modal domain restriction

We saw in the developmental section that modal domains, and domains of possibilities more generally, are constrained by normality restrictions. Those restrictions come in by default and are not part of modal cognition *per se*. They affect all types of reasoning where possibilities are under consideration, including fictional possibilities, and do not have to be indicated explicitly. Interestingly, the modal vocabularies of natural languages include items that explicitly direct us *away* from default normality restrictions (Rubinstein, 2012, 2014; Phillips & Cushman, 2017; Phillips, et al., 2019; von Fintel & Iatridou, 2023). (15) is one of Rubinstein's examples.

(15) [Preparing a company's tax report.]

a. We **have to** report all of our revenue.

b. We **should** report all of our revenue.

Rubinstein (2014, 538).

15(a), which has the all-purpose modal 'have to', relies on default normality restrictions for the possibilities considered, which are presumed to be shared without need for prior negotiation. In this particular case, the possibilities considered all conform to current tax laws. In contrast, as Rubinstein observes, the 'should'-claim in 15(b) conveys that tax evasion is among the possibilities considered, albeit judged to be non-optimal. On this view, both 'have to' and

(i) She promised yesterday that she would see me before she **left** next month.

The reason for the past tense on 'left' is that the 'before'-clause is in the scope of the past tense marker in the main clause.

‘should’ are necessity modals, but they differ in the way their domains are restricted: All-purpose ‘have to’ relies on default normality restrictions. It places no lexical restrictions on modal domains. In contrast ‘should’ lexically signals reliance on modal domain restrictions that deviate from the default ones.[13]

The existential modals ‘could’ and ‘might’ differ in a similar way: all-purpose ‘could’ relies on default normality restrictions, which include constraints relating to what is morally acceptable, whereas the lexically more specialized ‘might’ signals a departure from the default. In this case, departing from the default means that considerations of what is normally acceptable on moral grounds do no longer play a role. The Evil Ship Captain scenario illustrates.

(16) Evil Ship Captain

> FitzRoy is a notoriously ruthless pirate who decided to pose as a captain of a ship in order to steal several expensive sculptures that a museum needs to transport across the sea. After he posed as an ordinary ship captain who was taking some passengers across the sea, FitzRoy got the job.
>
> While sailing on the sea, a large storm came upon FitzRoy and his small ship. As the waves began to grow larger, FitzRoy realized that his small vessel was too heavy and the ship would flood if he didn’t make it lighter. The only things on FitzRoy's small boat were the expensive art sculptures that he was stealing and the passengers he was transporting. He knew he had to throw something overboard to keep the ship from capsizing.

In the context of Evil Ship Captain, compare 17(a) and (b).

[13] See also Yalcin (2016).

(17) a. FitzRoy **could** throw the sculptures overboard.

b. FitzRoy **might** throw the sculptures overboard.

The possibilities we consider for ‘could’ in 17(a) are influenced by what would be morally acceptable. This should not be surprising. Moral considerations are independently known to provide a default constraint on the possibilities considered and the way they are ranked. This has been separately observed for the possibilities considered in causal reasoning (e.g., Alicke, 2000; Knobe & Fraser, 2008; Kominsky & Phillips, 2019; Kominsky, Phillips, Gerstenberg, Lagnado, &. Knobe, 2015), the possibilities considered in making judgments of force and freedom (e.g., Phillips & Knobe, 2009; Phillips & Young, 2011; Bernhard, LeBaron, & Phillips, 2022), the possibilities entertained in counterfactual reasoning (e.g., McCloy & Byrne, 2000; N’gbala & Branscombe, 1995; Byrne & Timmons, 2018), the possibilities entertained when asking explicitly about what is possible (e.g., Phillips & Cushman, 2017; Shtulman & Phillips, 2018), and in many other cases as well (e.g., Pettit & Knobe, 2009; Phillips, Lugri, & Knobe, 2015). In the Evil Ship Captain scenario, FitzRoy throwing the sculptures overboard is clearly among the highest ranked possibilities on a moral scale, and we thus agree with 17(a). For ‘might’, on the other hand, normal moral considerations are blocked from influencing the possibilities considered. It is this blocking of the default that seems to be lexically signaled by using ‘might’, rather than ‘could’. With default moral considerations excluded, the possibilities considered for ‘might’ in this context seem to be restricted and ranked in a way that reflects what is likely to happen. Given FitzRoy’s evil nature, the highest ranked possibilities for ‘might’ are less likely to include possibilities in which the sculptures, rather than the passengers, are thrown overboard, hence our comparative disagreement with 17(b). An opposing pattern can be found when comparing 18(a) and (b).

(18) a. FitzRoy **could** throw the passengers overboard.

b. FitzRoy **might** throw the passengers overboard.

Here, the reliance on default normality constraints of 'could' but not 'might', should lead us to agree with 18(b) more than with 18(a).

We confirmed these predictions experimentally. Participants were first asked to read the Evil Ship Captain context, and then were asked to rate their agreement with all four modal claims 17(a), (b) and 18(a), (b) in random order. For simplicity we focus here on participants' first responses, though similar patterns are found when analyzing all of participants' responses. Overall, we found the predicted interaction between the modal term used ('could' vs. 'might') and whether the modal claim concerned throwing passengers or sculptures overboard**,** $F(1,185) = 54.730$, $p < .001$, $\eta^2_p = 0.228$ (Fig. 11). Specifically, participants more agreed that FitzRoy *could* throw the sculptures overboard ($M = 92.17$; $SD = 14.29$) than they agreed that he *might* ($M = 46.60$; $SD = 30.67$), $t(56.28) = 8.828$, $p < .001$, $d = 1.948$. And, in contrast, they less agreed that FitzRoy *could* throw the passengers overboard ($M = 65.13$; $SD = 37.04$) than they agreed that he *might* ($M = 79.89$; $SD = 22.97$), $t(90.80) = -2.422$, $p = .017$, $d = 0.469$.[14]

[14] Additional experimental details, including stimuli, materials, data, code, sample sizes, participant exclusions, and a longer explication of the results can be found in the supplement to our paper: https://doi.org/10.7910/DVN/KUWNYK. This study was not preregistered.

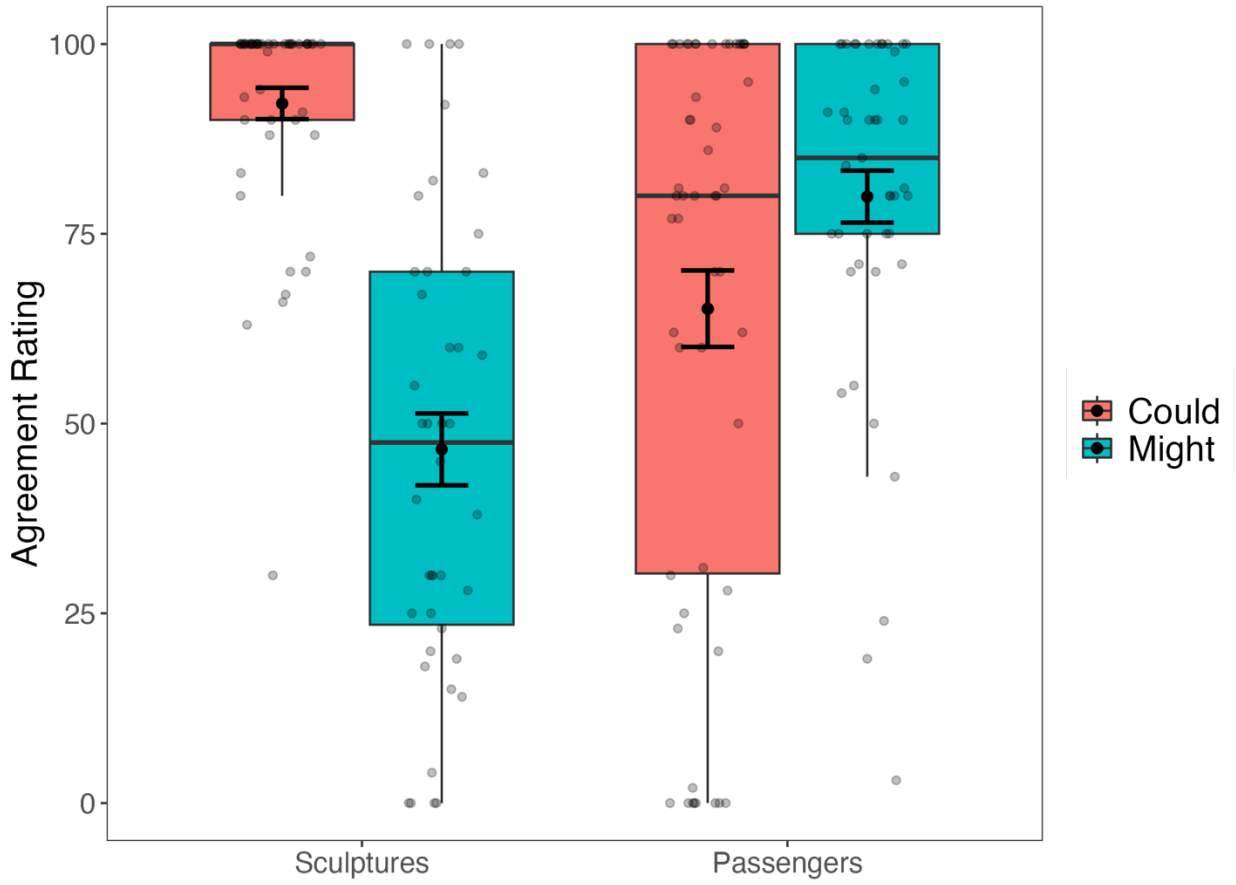


Fig. 11. Boxplots of participants' agreement the existential modal claims when they concerned ratings FitzRoy throwing the passengers overboard (left boxes and points) or the sculptures overboard (right boxes and points), both when the existential modal used was 'could' (red boxes) and when it was 'might' (blue boxes). The colored boxes depict the middle two quartiles of responses, the horizontal line depicts the median response, and the small grey dots depict individual participant responses. Thick black dots depict mean agreement and error bars represent +/- 1 SEM.

Further evidence for taking lexically specialized modal vocabulary items to indicate departures from default normality restrictions can be found in studies that require participants to make a range of different modal judgments either very quickly or after taking a moment before

responding (Phillips & Cushman, 2017, Acierno, Mischel, & Phillips, 2022). If normality-based domain restriction is a default which is only moved away from with additional specific lexical information, then we should expect that the differences between the kinds of domain restriction indicated by different modal terms (e.g. ‘could’ vs. ‘might’) would require some additional, non-default, processing. Accordingly, we should then also expect that if people were not allowed enough time to complete that processing before having to respond, their judgments concerning these two different modal terms will begin to look more similar to one another, as they would all be forced to rely on default, normality-based, domain restriction. This prediction of our account of default normality restrictions is also borne out. Phillips and Cushman (2017) presented participants with different background contexts and asked them to make truth-value judgments of modal statements that concerned various events occurring in those contexts. Participants were either forced to respond quickly or were asked to reflect before responding. Phillips and Cushman then computed the similarity between (among others) modal judgments with ‘could’ vs ‘might’, both when participants were responding quickly and when they were responding slowly. This analysis revealed that these modal judgments became more similar to one another when participants were forced to answer quickly but were significantly less similar when participants were allowed additional time before responding (Phillips & Cushman, 2017). Studies focusing on the processing required when making different modal judgments thus provide support for a view on which normality-based domain restriction serves as a default that can be moved away from with additional processing when the lexical information of a more specialized modal requires it.

Summary and conclusion: How natural languages (de)compose modal thought

In this section, we have documented – within the limits of an article – that natural languages compositionally construct modal thoughts from the same components that we identified by

looking at which non-linguistic tasks can be passed by children at various stages of development. Natural languages represent modal anchors via the argument structure of modal expressions, they use tense to temporally locate modal anchors, and aspect to mark the temporal orientation of the possible extensions of the anchor situation that are being considered. We also saw with examples from English that modal vocabularies might include items that signal that the assumed modal restrictions deviate from the default ones. Surprisingly, English doesn't seem to have dedicated morphological means to express counterfactuality. It only marks two necessary – but together not sufficient – components: pastness of the modal anchor and future orientation.

It would be naïve, of course, to expect a completely transparent and predictable encoding of the building blocks of modal thought across languages. We should not underestimate our ability to import information that was not explicitly mentioned into the process of meaning composition, nor should we forget that there often is more than one grammatical way to assemble a particular meaning from available semantic building blocks.

Fitting modal thought and talk together

We have outlined what we take to be the main components of actuality-directed modal thought, how these pieces can be seen in cognitive development, and how they may be expressed in natural languages. We now want to consider one further question: when we utter (or understand) a sentence involving a modal expression, how do the linguistic components we've illustrated relate to the non-linguistic components of modal thought we began with? How do modal thought and talk fit together? The very beginning of an answer to this question, we think, can be found in some routine observations. It is probably not by mere accident that similar

component pieces can be posited to explain variation in the development of modal thought and the linguistic expression of modal thought across languages. Our non-linguistic modal thoughts can best be shared with others when we can compress them into a linguistic representation that reflects their structure. To do this optimally, linguistic representations of modality would be expected to encode the component pieces that are required for other minds to generate a representation of possibilities that corresponds to our own. We cannot share our non-linguistic modal thoughts with others directly, but we may have mechanisms for recreating them. Modal language provides a way of encoding the necessary ingredients of modal thought in a linguistic form that allows them to be shared.

With these observations in hand, we can now formulate a more specific question: Which ingredient pieces do linguistic representations of modality encode, and which pieces are left to non-linguistic cognition for sorting out? The account we've offered provides some illustrations where the division of labor begins to become clearer. Linguistic representations of modality do not encode all of the relevant assumptions required to restrict the domain of possibilities; some of this work is left to non-linguistic cognition. Linguistic representations might instead mark deviations from default normality-based domain restrictions. Also, modal expressions may explicitly encode the modal anchor, as with adjectives like *fragile*, and they may or may not mark the time of the anchor, or explicitly indicate past or present orientation of possible anchor extensions. There are also cases where the labor is shared: picking out the relevant anchor situation for English modal auxiliaries like *could* requires collaboration between linguistic and non-linguistic cognition.

Natural languages vary in what is and is not explicitly encoded. For example, some languages encode the force of a modal claim (e.g. existential vs. universal), while others do not (Rullmann, et al., 2008; Deal, 2011). Some languages have lexically specified deontic modals (Rullmann, et al., 2008), others – like English – do not. As we have seen, English does not morphologically

mark counterfactuality, nor does it have modals that require their anchors to be epistemic states. There might not be any straightforward 'one fits all' recipe for how modal thought and talk fit together in natural languages. There doesn't have to be: there is usually more than one way to compositionally put together a given modal thought and there is usually also more than one way to encode enough of the component pieces to allow for non-linguistic cognition to infer the remaining ones within a tolerable range of unresolved indeterminacy. Each language offers a blueprint for how this can be done.

Towards a unified theoretical framework

Our discussion so far has accumulated support for our particular view of modal thought from two distinct empirical domains: cognitive development on the one hand and language typology and acquisition on the other. We have made a case for a new way of thinking about modal thought by singling out modal thought that is actuality-directed and by positing a distinctive cognitive mechanism responsible for it: factual domain projection from an actual anchor. In this concluding section, we will suggest that the mechanism of factual domain projection is the bridge that we think can pave the way towards a theory of modal thought where the results of different disciplines each have their place, their potential relevance to each other is understood, and researchers working with different methodologies can speak a common language. In what follows we will thus connect our own proposal to existing theoretical work in philosophy, logic, computer science, and cognitive psychology.

Connection to formal frameworks from philosophy, logic, and semantics

The formal semantic investigation of modal thought originated in modal logic. Kripke's (1963) semantics for existing axiomatic systems of modal logic provided not only a semantic characterization of some valid principles of modal reasoning, but its possible worlds framework was soon recognized as foundational for the semantics of natural languages more generally: Knowledge of meaning implies the ability to connect sentences with the sets of possible worlds where they are true (Lewis 1968, 1970; Montague 1970). A shared framework based on possible worlds is responsible for the fact that linguists, philosophers, and logicians interested in notions related to possibility and necessity largely already speak a common language today.

The possible worlds framework is based on the truism that the world we live in could be different in countless ways. We can think of each of those ways as instantiated by a possible world. If possible worlds are the same kinds of thing as the actual world, they are specified down to every detail. As a consequence, they can't possibly be represented by the minds of us mere mortals. However, it would be wrong to think that those same mortals wouldn't be capable of thinking of meanings in terms of an infinite plenitude of possible worlds. After all, they are perfectly capable of reasoning about the infinities contained in the continuum of the real numbers, for example. Still, there is an important question about how human minds might go about representing possibly infinite domains of possibilities. The theory of mental models of Johnson-Laird and colleagues offers an answer, which we will return to shortly. More generally, a possibly infinite set of possible worlds can be represented via a representable property that all worlds in the set share, even though they may otherwise differ in every which way. Suitable modes of representation might include pictures, sentences, architectural models, geological maps, scatter plots, mental models, and what have you.

Kripke's semantics for modal logic was the starting point for a semantics of modal expressions in natural languages like the English auxiliaries 'must', 'might', or 'can'. In Kripke's semantics modal operators are existential or universal quantifiers over domains of possible worlds. The domains of those quantifiers in a given world w are sets of worlds that are related to w via a binary relation R, the 'accessibility relation'. A proposition is possible in a world w if it is true in some world related to w by R. It is necessary if it is true in every world related to w by R. Different systems of modal logic can be characterized by placing different formal restrictions on R. R may be required to be reflexive or transitive, for example. Crucially, there is no commitment to a particular value for R. This is so because, when studying principles of modal reasoning, we are interested in inferences that are valid regardless of a particular choice of R (that is, regardless of a particular choice of modal domain), as long as R satisfies certain formal constraints.

Unlike logicians, formal semanticists working on natural language are not merely interested in characterizing modal inferences that are valid regardless of the choice of modal domains. Since knowledge of sentential meanings includes knowledge of truth-conditions, their primary concern are the contributions of linguistic expressions to truth-conditions. To illustrate, a theory that only characterizes valid modal reasoning may tell you that if there must be a pot on the shelf it follows that there might be one there, but it has nothing to say about what the actual world has to look like for your friend's claim that you can ship a particular pot in a particular box to be true. This last kind of question is not only crucial for natural language semantics, it is also the crucial question for our project, and this is where factual domain projection enters the scene.

For natural language semantics, knowledge of truth-conditions is the basic notion. Knowledge of whether a given sentence is a tautology or a contradiction, or whether a particular inference is

logically valid is derivative. To illustrate, if you know in what kind of possible situations a sentence is true, you know whether it is true in all possible situations or in no possible situation, for example, hence whether it is a tautology or a contradiction. You are also in a position to figure out logical relations between sentences, since these relations all correspond to relations between the sets of possible situations expressed by those sentences. The logical consequence relation corresponds to set inclusion, for example.

The fact that language users know the truth-conditions for sentences with modals in their language means that they know how modal domains are set up. This calls for an account of how the domains of modals are established by actual people in actual situations. Linguists don't give us such an account. Our proposal that modal domains for actuality-directed modal thought are projected by considering possible extensions of an actual anchor situation forms the basis of such an account that can then be fleshed out further within cognitive psychology or computer science by investigating possible algorithms that realize such a function, as we discuss below. Staying within formal semantics, the account requires richer formal models than those used in classical modal logic. Instead of a set of possible worlds, we need a set of possible situations linked by a part relation where worlds are maximal elements (see Kratzer, 2021 for an overview). In addition, our models have to include binary relations between situations corresponding to the match relation and to temporal relations (simultaneous, later, earlier). Against the background of such a model, our function f_{act} maps situations s to the set of situations that have as a part a situation s' that stands in the match relation to s. A proposition p is possible with respect to an anchor situation s just in case there is a situation in f_{act} (s) where p is true. A proposition p is necessary with respect to an anchor situation s just in case every situation in f_{act} (s) is part of a situation where p is true. Intuitively, this last condition says that a proposition p is necessary with respect to an anchor situation s just in case the facts in s guarantee the truth of p. If p isn't already true in s, this is a condition that is hard to satisfy, and

this is why normality constraints play such an important role in modal reasoning. Modal domains can be restricted further by temporal relations so as to only contain situations that are in the future or past of the anchor situation, or are simultaneous with it.

To illustrate in more detail how these formal frameworks combine with our proposal, here is a formula representing the truth conditions for a particular utterance of the sentence 'The glass could break' to be true in an actual context that provides the anchor situation s_0 and the normality restrictions N.[15]

(19) $\exists w\,(w \in f_{act}(s_0)\ \&\ N(w)\ \&\ \exists s\,(s \leq w\ \&\ s_0 \approx s\ \&\ \exists s'\,(s \preccurlyeq s'\ \&\ \mathit{break}(\mathit{the\ glass})(s'))))$

The formula is the result of a computation that assembles it step-by-step from the semantic values of the constituents of the sentence and contextual contributions. Here are the essential ingredients of the formula:

f_{act}	The function responsible for factual domain projection
N	Normality restrictions
$\leq$	The part-whole relation between situations
$\approx$	The match relation between situations
$\preccurlyeq$	The temporal precedence relation between situations, responsible for temporal orientation
s_0, s, s'	situations; s_0 is the anchor situation
w	worlds, maximal elements of $\leq$

The computation of truth-conditions for a sentence is modeled via expressions of a λ-calculus in formal approaches to natural language semantics. The computation may proceed from a more abstract representation as in 20(a), which instantiates the general schema in 20(b).

[15] For this illustration, we are neglecting the possibility that 'could' may also describe possibilities in the past, in which case it would be taken to be a past tense modal.

(20) a. (could (future (the glass break)))

b. (modal (aspect (clause)))

The semantic contributions of the three pieces represented in 20(b) would be as in (21), where both aspect and the modal contribute functions from propositions, characteristic functions of sets of possible situations.

(21) a. $\lambda s\ break(the\ glass)(s)$ clause

b. $\lambda p \lambda s\ \exists s'\ (s \preccurlyeq s' \ \&\ p(s'))$ aspect

c. $\lambda p\ \exists w\ (w \in f_{act}(s_0) \ \&\ N(w) \ \&\ \exists s\ (s \leq w \ \&\ s_0 \approx s \ \&\ p(s)))$ modal

The formula in (19) is the result of combining the components in 21(a) to (c) according to the general schema in 20(b), where the mode of combination is Functional Application in both cases.

(22) a. aspect (clause) = $\lambda p \lambda s\ \exists s'\ (s \preccurlyeq s' \ \&\ p(s'))\ (\lambda s\ break(the\ glass)(s)) =$

$\lambda s\ \exists s'\ (s \preccurlyeq s' \ \&\ break(the\ glass)(s'))$.

b. modal (aspect (clause)) =

$\lambda p \exists w(w \in f_{act}(s_0) \ \&\ N(w) \ \&\ \exists s\ (s \leq w \ \&\ s_0 \approx s \ \&\ p(s)))$

$(\lambda s\ \exists s'\ (s \preccurlyeq s' \ \&\ break(the\ glass)(s'))) =$ (19).

(19) says that there is a world w in the modal domain projected by f_{act} from the anchor situation s_0 and constrained by normality restrictions N with the following properties: There are parts s and s' of w such that s is a match of s_0, s temporally precedes s', and the glass breaks in s'.

To summarize this subsection, our proposed core ability for actuality-directed modal thought can be formally modelled within a situation-based semantic framework as a function that maps situations s to the set of situations that have a match of *s* as a part. This function delivers modal domains, and in so doing, it provides the crucial bridge that connects theoretical frameworks from modal logic and formal semantics of natural languages to our earlier account of the development of modal thought and its expression in natural languages.

Connection to formal frameworks from cognitive psychology and computer science

A number of prominent approaches in cognitive psychology and computer science have built on and extended logical frameworks involved in reasoning over possibilities. However, despite the points of formal continuity between this work and the work discussed in the prior section, the two approaches also differ critically in what they aim to capture. As we noted before, the aim of the work in the prior section was to capture certain truths: the conditions that would make a given sentence true, the semantic characterization of valid modal reasoning, and so on. These proposals can be understood as telling us something at Marr's 'computational' level—they are characterizations of the form of the problem that humans must solve when thinking about possibilities. By contrast, the work that we'll be discussing from psychology and computer science typically aims to characterize, at various levels of abstraction, the way that minds may actually go about reasoning about possibilities—they are proposals more naturally thought of at Marr's 'representational' or 'algorithmic' level. Accordingly, we think of these two different bodies of work not as competing proposals, but as broadly consistent ways of solving two related puzzles. The first tells us about the boundaries or rules of the game; the second provides a way of actually playing it. A benefit of our proposal based on factual domain projection is that it makes it easier to see the connection between the two bodies of work; it provides a bridge. In

what follows, we'll illustrate these relationships with two examples of two formal frameworks: mental models and recent approaches to probabilistic modeling in cognitive science.

Mental models

One research program has focused on understanding the 'mental models' that people employ when reasoning. Developed by Philip Johnson-Laird and colleagues (see Johnson-Laird & Ragni, 2019 for a recent overview), the mental models theory shares with standard possible worlds semantics the assumption that sentence meanings are representations of the possibilities where the sentence is true. To illustrate, consider (19).

(19) Chris has a dog or a cat, or both a dog and a cat.

In standard possible worlds semantics, (19) describes the set of possible worlds where Chris has a dog or a cat or both a dog and a cat. Relying on experimental evidence, the mental model theory makes a claim about how the possibilities described by (19) are mentally represented. Experimental evidence suggests that they are represented as the three mutually incompatible possibilities *A*, *B*, and *C*, where, from our perspective, *A* would correspond to the set of worlds where Chris has a dog, but no cat, *B* would correspond to the set of worlds where she has a cat, but no dog, and *C* would correspond to the set of worlds where she has both a dog and a cat. The possibilities posited by the mental models theory thus correspond to the cells of a partition of the sets of possible worlds where (19) is true. The example shows that mental model theory goes beyond standard possible worlds semantics by positing more fine-grained meanings in terms of sets of mutually incompatible possibilities.[16] Critically, the impetus for this extension

[16] A recent semantic theory that comes close in spirit to mental models theory is Inquisitive Semantics (see Ciardelli & Roelofsen, 2018). In Inquisitive Semantics, sentence meanings are also construed, not as sets of possible worlds, but as sets of sets of possible worlds, which, in that theory, don't have to be mutually incompatible,

beyond standard possible world semantics is that mental models theory aims to capturing how people actually represent and reason about possibilities. Thus, a major target for this theory is to recapture successful and valid reasoning about possibilities, but also why people sometimes make errors, *e.g.*, denying the antecedent or confirming the consequent (Johnson-Laird, 2001; Johnson-Laird, 2010).

Turning to the relationship to our proposal, it is important to note that the deductive reasoning targeted by mental models theory is importantly different from the actuality-directed modal thought we have sought to explain. While both kinds of thought clearly relate to possibilities, the theory of mental models is primarily concerned with the question of how possibilities are represented when people engage in reasoning about logical relationships between sentences, including sentences with modals. However, it does not concern the kind of actuality-directed modal thought we engage in when we think about what we could pack in a box, guess that it might have rained when we see a puddle of water, or wonder what would have happened if Susie had taken off her shoes before walking into the kitchen. These are the kinds of modal thoughts we have sought to offer an account of, and so we see the two accounts as complementary and entirely compatible with each other. They each seek to solve a different part of a broader puzzle concerning the role of possibilities in thought.

Probabilistic models

A different formal approach that has recently been pursued in work in cognitive psychology and computer science is to use probabilistic models to capture aspects of cognition that involve generating or reasoning over possibilities. These models enrich the logic-based approaches described above in two primary ways. The first is that possibilities are mapped to a continuous

however. See Koralus & Mascarenhas (2013) for work on the connection between mental models and inquisitive semantics.

measure (a real number between 0 and 1), which represents how 'possible' each possibility is treated as being. The second is that probabilistic models often rely on an enrichment of the unstructured nature of possible worlds, adding *e.g.*, causal structure to the possibilities reasoned over. Recent proposals using probabilistic models have had remarkable success capturing predictions and inferences about physical events (Smith, Battaglia, & Tenenbaum, 2023; Smith, Mei, Yao, Wu, Spelke, Tenenbaum, & Ullman, 2019; Zhou, Smith, Tenenbaum, & Gerstenberg, 2023), decision making (Halpern, 2017; Lieder & Griffiths, 2020; Vul, Goodman, Griffiths, & Tenenbaum, 2014), causal judgments (Pearl, 2009, Gerstenberg, et al., 2021; Glymour, 2001), pragmatic reasoning (Degen, 2023; Frank & Goodman, 2012; Franke & Jäger), and many among other things as well (for reviews, see, Chater, Tenenbaum, & Yuille, 2006; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010).

As we said at the outset, the core of the connection between our proposal and research projects that employ probabilistic models concerns factual domain projection. Probabilistic models of modal thought involve some algorithmic specification of a function that returns a domain of possibilities. That is, they typically involve considering what is possible (in probabilistic terms) given some particular state of affairs (and then go on to specify how one might do some further reasoning involving those possibilities). Getting a sense for how this works and how it connects to factual domain projection will be helpful with a concrete example.

Consider, for example, work that models people's physical predictions in cases of uncertainty, such as whether or how a tower of blocks will fall under various conditions or interventions (*e.g.*, Battaglia, Hamrick, & Tenenbaum, 2013; Zhou, Smith, Tenenbaum, & Gerstenberg, 2023). Predictions such as this are straightforward cases of factual domain projection—one takes the actual blocks at the current time in their current position as the modal anchor and returns a set of possibilities that involve that exact set of blocks now extending into the future. At our level of abstraction, the possibilities of interest (the situations in the modal domain) are the set of

situations {s: s ∈ $f_{act}(s_0)$ & N(s)}, where s_0 is the anchor situation, and N a condition corresponding to the relevant normality constraints.

Our own proposal tells us something about the *kind* of possibilities that must be considered but does not tell us precisely how one would go about actually generating such a set in a computationally efficient manner. Helpfully, algorithmic ways of implementing this kind of factual domain projection have been beautifully spelled out using probabilistic models. For example, Battaglia et al. (2013) model predictions of whether a tower of blocks will fall with a model that uses a probability distribution over the physical representation of an initial state of the scene at t_0 and a probability distribution of the latent physical forces in the scene, and then simulates a sequence of future states by recursively applying a set of deterministic physical dynamics to the initial state (using a physics engine). This algorithmic process is used to generate a set of states at time t_n which can then be queried for whether or how the blocks fell at that (future) time. Critically, the set of states at time t_n returned provide a modal domain, or set of situations that are the future extensions of exact matches of the anchor situation, or initial state of the blocks. It is worth noting explicitly that not only does this probabilistic model instantiate a form of factual domain projection, but the modal domain returned is constrained by *normality*: this is guaranteed by the fact that the physics engine used respects the laws of physics and the initial state representations are rationally inferred from the observed state of the scene.

Or, to take a different example, another widespread use of probabilistic models has been to capture how people make causal judgments by modeling how one can reason counterfactually about what would have happened if things had been different (Pearl, 2009, Gerstenberg, et al., 2021; Glymour, 2001). In our discussion of counterfactual thought and language, we proposed that one type of counterfactual modal thought involves selecting a past anchor and then projecting a forward-oriented domain of possibilities (possibilities that involve an exact match of

the past situation but extend forward in time). While our proposal again tells you something about the kinds of possibilities one should be evaluating, it says very little about how exactly to go about accomplishing this. Helpfully, precise algorithmic proposals have again been offered for how such cognition may work in specific cases involving causal judgments of physical events and human agents (Gerstenberg, et al., 2021; Icard, Kominsky, Knobe, 2017; Quillien & Lucas, 2023). While these models differ in their details, what they all share is that they involve ways of generating a set of counterfactual possibilities by probabilistically simulating what else could have happened in some alternative future. For example, in work on judging the cause of winning a lottery that involves drawing from urns (Morris, Phillips, Icard, Knobe, Gerstenberg, & Cushman, 2018; Quillien & Lucas, 2023), these models involve going back to the time prior the draw and simulating alternative possible draws from the urns. They then go on to predict people's causal judgments by assuming that they are comparing what actually occurred to what else would have occurred in this set of counterfactual possibilities (*i.e.*, the modal domain). While all of these models share this underlying structure, they primarily differ in how they predict causal judgments by focusing on different aspects of the relationships between the events that occur in the probabilistically generated counterfactual possibilities. And once again, it is worth noting that in line with our more abstract proposal, all of these different models involve some way of ensuring that the set of counterfactual possibilities are constrained by normality, whether indirectly (*e.g.*, Gerstenberg, et al., 2021) or directly (*e.g.*, Icard et al., 2017).

The promise of bridges

A major theoretical benefit of our proposal is that it allows us to see how different bodies of research on possibilities are connected to one another (from developmental and comparative psychology, to formal semantics, to modal logic, to probabilistic models). From this more abstract perspective, it also becomes easier to see how discoveries made in one research

program can and should inform other research programs. As an illustration, consider the work that points out that forward inferences in causal Bayes nets—predictions of some effect given some cause, P(e|c)—should be (and empirically are) easier to make than backward inferences—diagnoses of causes given some effect, P(c|e)—because of the possibility of alternative competitive causes that may explain that same effect (Meder, Mayrhofer, & Waldmann, 2014; Waldmann, 2000; Waldmann & Holyoak, 1992). This distinction between predictive and diagnostic reasoning aligns with differences in the temporal orientation of factual domain projection. Prediction involves forward-oriented domain projection; diagnosis involves backward-oriented domain projection (which we illustrated in one of our opening examples of reasoning that it must have rained given the puddle on the ground). Thus, an intriguing possibility is that the discovery about differences in predictive and diagnostic reasoning may point to a more general difference in the difficulty of forward- vs. backward-oriented domain projection. This is an interesting avenue for future research, perhaps especially for those working on modal cognition in development.

To end with a second illustration, our proposal makes a major set of predictions around the distinction between epistemic and non-epistemic modal anchors and uses developmental research to demonstrate the importance of this distinction. As far as we know, this crucial distinction has not been incorporated in other work that empirically investigates or models modal thought, though doing so is likely quite important. Deciding what is possible given some piece of the outside world is importantly different from deciding what is possible given your representation of the outside world, and only the latter requires you to have a model of your own cognition (and ability to meta-represent one's own epistemic state). For example, consider computational work that builds artificial agents who make inferences and predictions about the environment they are in. Typically, these artificial agents build a model of the environment directly, and do not have the capacity to model their own model of the environment (De Freitas,

Uğuralp, Oğuz-Uğuralp, Paul, Tenenbaum, & Ullman, 2023; Paul, Ullman, De Freitas, & Tenenbaum, 2023). This idea of generating a set of possibilities based on one's own epistemic state can be usefully applied not only in computer science, but also in formal semantics, modal logic, probabilistic models, and comparative and developmental psychology.

Conclusion

We've drawn attention to modal thought that is actuality-directed and we have offered a way of decomposing it into its component pieces. Actuality-directed modal thought is not a monolithic capacity that would be expected to emerge fully formed. It is a complex capacity built from simpler parts that can be productively combined in different ways to allow for an increasingly impressive range of abilities. We have supported our proposed decomposition with evidence from the development of modal thought and from the way it is expressed in natural languages. Last not least, we have shown how our proposal fits into current theorizing in modal logic, formal semantics for natural languages, cognitive psychology, and computer science.

Our proposed decomposition provides a starting point for understanding how natural minds project possibilities from parts of the world they live in. As researchers across the cognitive sciences continue to study modal thought, we suspect that new discoveries will require aspects of our proposal to be refined or abandoned altogether. But our hope is that the proposal will be a first step towards a unified theory of modal thought that allows developmental psychologists to make claims more easily understood and modeled by logicians, for example, or comparative cognition researchers to test ideas originally formulated by philosophers and semanticists.

References

Acierno, J, Mischel, S., & Phillips, J. (2022). Moral judgements reflect default representations of possibility. *Phil. Trans. R. Soc. B*, *377*: 20210341.

Aikhenvald, A. (2004). *Evidentiality*. Oxford University Press.

Aikhenvald, A. (Ed.). (2018). *The Oxford Handbook of Evidentiality*. Oxford University Press.

Alderete, S., & Xu, F. (2023). Three-year-old children's reasoning about possibilities. *Cognition*, 237, 105472.

Alicke, M. D. (2000). Culpable control and the psychology of blame. Psychological Bulletin, 126(4), 556.

Arregui, A. (2009). On similarity in counterfactuals. Linguistics & Philosophy, 32, 245-278.

Arregui, A. (2021). Counterfactuals: 'If kangaroos had no tails ...' In D. Gutzmann, L. Matthewson, C. Meier, H. Rullmann, & T. E. Zimmermann (eds.): *The Wiley Blackwell Companion to Semantics*. John Wiley & Sons, Inc. DOI: 10.1002/9781118788516.sem007.

Baillargeon, R. (1987). Object permanence in 3½-and 4½-month-old infants. *Developmental Psychology*, 23(5), 655.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327-18332.

Beck, S.R., Robinson, E.J., Carroll. D.J., & Apperly, I.A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, 77(2):413-26.

Beran, M. J., Perner, J., & Proust, J. (Eds.). (2012). *Foundations of metacognition*. Oxford University Press.

Bernhard, R. M., LeBaron, H., & Phillips, J. (2022). It's not what you did, it's what you could have done. *Cognition*, 228, 105222.

Bell, V. A., & Johnson-Laird, P. N. (1998). A model theory of modal reasoning. *Cognitive Science*, 22(1), 25-51.

Bhatt, R. (1999). Covert Modality in Non-Finite Contexts. University of Pennsylvania PhD dissertation.

Browne, C. A., & Woolley, J. D. (2004). Preschoolers' magical explanations for violations of physical, social, and mental laws. *Journal of Cognition and Development*, 5(2), 239-260.

Byrne, R. M. (2005). *The Rational Imagination: How people create alternatives to reality*. MIT press.

Byrne, R. M., & Timmons, S. (2018). Moral hindsight for good actions and the effects of imagined alternatives to reality. *Cognition*, 178, 82-91.

Carey, S., Leahy, B., Redshaw, J., & Suddendorf, T. (2020). Could it be so? The cognitive science of possibility. *Trends in Cognitive Sciences*, 24(1), 3-4.

Cesana-Arlotti, N., Téglás, E., & Bonatti, L. L. (2012). The probable and the possible at 12 months: Intuitive reasoning about the uncertain future. *Advances in child development and behavior*, 43, 1-25.

Cesana-Arlotti, N., Kovács, Á. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature Communications*, 11(1), 5999.

Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science*, 359(6381), 1263-1266.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7), 287-291.

Chen, S., Hohaus, V., Laturnus, R., Louie, M., Matthewson, L., Rullmann, H., Simchen, O., Turner, C. K., & Vander Klok J. (2017). Past possibility crosslinguistically: Evidence from twelve languages. Arregui, A., Rivero M. L., & Salanova, A. (Eds.). (2017). *Modality across Syntactic Categories*. Oxford University Press, 235-287.

Chernyak, N., Kushnir, T., Sullivan, K. M., & Wang, Q. (2013). A comparison of American and Nepalese children's concepts of freedom of choice and social constraint. *Cognitive Science*, 37(7), 1343-1355.

Ciardelli, I. & Roelofsen, F. (2018). *Inquisitive Semantics*. Oxford University Press.

Condoravdi, C. (2002): Temporal Interpretation of modals: Modals for the present and for the past. In D. Beaver, S. Kaufmann, B. Clark and L. Casillas (eds.). *The Construction of Meaning*, 59–88. CSLI Publications.

Cournane, A. (2020). Learning modals: A grammatical perspective. *Language and Linguistics Compass*, *14*(10), 1-22.

Cournane, A. (2021). Revisiting the epistemic gap: It's not the thought that counts. *Language Acquisition*, 28(3), 215-240.

De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, *51*(12), 2401-2414.

De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., Paul, L. A., Tenenbaum, J., & Ullman, T. D. (2023). Self-orienting in human and machine learning. *Nature Human Behaviour*, 1-14.

Deal, A. R. (2011). Modals without scales. *Language*, 87(3), 559-585.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, 9, 519-540.

Engelmann, J.M., Völter, C.J., O'Madagain, C., Proft, M., Haun, D.B.M, Rakoczy, H., & Herrmann, E. (2021). Chimpanzees consider alternative possibilities, *Current Biology*, 31(20): R1377-R1378.

Engelmann, J.M., Völter, C.J., Goddu, M. K., Call, J., Rakoczy, H., & Herrmann, E. (2023). Chimpanzees prepare for alternative possible outcomes. *Biol. Lett*. 19: 20230179. https://doi.org/10.1098/rsbl.2023.0179

von Fintel, K., & Iatridou, S. (2023). Prolegomena to a theory of X-marking. *Linguistics & Philosophy*. https://doi.org/10.1007/s10988-023-09390-5.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998-998.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1), 3-44.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25–42.

Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences*, *3*(11), 410-419.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936.

Glymour, C. N. (2001). The mind's arrows: Bayes nets and graphical causal models in psychology. MIT press.

Goddu, M. K., Sullivan, J. N., & Walker, C. M. (2021). Toddlers learn and flexibly apply multiple possibilities. *Child Development*, *92*(6), 2244-2251.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37(5), 620.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357-364.

Gweon, H., & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, 332(6037), 1524-1524.

Hackl, Martin. 1998. On the semantics of ability attributions. Manuscript, MIT.

Hacquard (2006): *Aspects of Modality*. MIT PhD dissertation.

Hacquard, Valentine. (2010). On the event-relativity of modal auxiliaries. *Natural Language Semantics* 18, 79-114.

Halpern, J. Y. (2017). Reasoning about uncertainty. MIT press.

Harris, P. (2022a). Young children share imagined possibilities: evidence for an early-emerging human competence. *Philosophical Transactions of the Royal Society B,* 377: 20220022. https://doi.org/10.1098/rstb.2022.0022

Harris, P. L. (2022b). Children's Imagination. Cambridge University Press.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, 61(3), 233-259.

Harris, P. L., Kavanaugh, R. D., Wellman, H. M., & Hickling, A. K. (1993). Young children's understanding of pretense. *Monographs of the Society for Research in Child Development*, i-107.

Hecht, E. & Phillips, J. (2022). Domain-general nodal spaces play a foundational role throughout high-level cognition. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.*

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, 106(11), 587-612.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, 161, 80-93.

Johnson, K. (2022). Time and evidence in the graded tense system of Mvskoke (Creek). *Natural Language Semantics* 30,155-183.

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in cognitive sciences*, *5*(10), 434-442.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43), 18243-18250.

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition*, *193*, 103950.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2), 99-134.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93(2), 136.

Kalish, C. (1998). Reasons and causes: Children's understanding of conformity to social rules and physical laws. *Child Development*, 69(3), 706-720.

Kloo, D., Rohwer, M., & Perner, J. (2017). Direct and indirect admission of ignorance by children. *Journal of Experimental Child Psychology*, 159, 279-295.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psychology*, 2, 441-8.

Kominsky, J. F., Gerstenberg, T., Pelz, M., Sheskin, M., Singmann, H., Schulz, L., & Keil, F. C. (2021). The trajectory of counterfactual simulation in development. *Developmental Psychology*, 57(2), 253.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137, 196-209.

Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives*, 27, 312-365.

Kratzer, A. (1981). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, 10(2), 201-216.

Kratzer, A. (1989). An investigation of the lumps of thought. *Linguistics and Philosophy*, 607-653.

Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives* (Vol. 36). Oxford University Press.

Kratzer, A. (2013). Modality for the 21st Century. In S. R. Anderson, J. Moeschler, F. Reboul (eds.): *The Language Cognition Interface / L'Interface Language-Cognition*. Geneva-Paris (Librairie Droz). 179-199.

Kratzer, A. (2021). Situations in natural language semantics. In E. Zalta (ed.): Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/situations-semantics/.

Kripke, S. (1963). Semantical Analysis of Modal Logic I. Normal Modal Propositional Calculi. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 9, 67-96.

Lane, J. D., Ronfard, S., Francioli, S. P., & Harris, P. L. (2016). Children's imagination and belief: Prone to flights of fancy or grounded in reality?. *Cognition*, 152, 127-140.

Leahy, B. P., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, *24*(1), 65-78.

Leahy, B., Rafetseder, E., & Perner, J. (2014). Basic conditional reasoning: How children mimic counterfactual reasoning. *Studia Logica*, 102(4), 793-810.

Leahy, B., Huemer, M., Steele, M., Alderete, S., & Carey, S. (2022). Minimal representations of possibility at age 3. *Proceedings of the National Academy of Sciences*, *119*(52), e2207499119.

Leahy, B., & Zalnieriunas, E. (2022). Might and might not: Children's conceptual development and the acquisition of modal verbs. In *Semantics and Linguistic Theory* (Vol. 31, pp. 426-445).

Lee, J. (2013). Temporal constraints on the meaning of evidentiality. *Natural Language Semantics*, 21(1), 1-41.

Lewis, D. K. (1968). *Convention*. Cambridge/Mass. (Harvard University Press).

Lewis, D. K. (1970). General Semantics. *Synthese*, 22(1/2), 18-67.

Lewis, D. K. (1973). Counterfactuals. Malden. Blackwell.

Lewis, D. K. (1974). Causation. *The Journal of Philosophy*, *70*(17), 556-567.

Lewis, D. K. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74(4), 549-567.

Lewis, D. K. (1997). Finkish dispositions. *The Philosophical Quarterly*, 47, 143-158.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Matthewson, Lisa. (2013). Gitksan Modals. *International Journal of American Linguistics,* 79, 349-394.

McCormack, T., & Hoerl, C. (2020). Children's future-oriented cognition. In *Advances in Child Development and Behavior,* 58, 215-253.

McCloy, R., & Byrne, R. M. (2000). Counterfactual thinking about controllable events. *Memory & Cognition*, 28(6), 1071-1078.

McHugh, D. (2023). Causation and Modality. Models and Meanings. Amsterdam. Institute for Logic, Language, and Computation.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, 121(3), 277.

Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, 154, 40-48.

Montague, R. (1970). Universal Grammar. *Theoria*, 36(3), 373-398.

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating options and choosing between them depend on distinct forms of value representation. *Psychological Science*, 32(11), 1731-1746.

Morris, A., Phillips, J. S., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Causal judgments approximate the effectiveness of future interventions. Unpublished Manuscript.

Murray, S. (2021a). Evidentiality, modality, and speech acts. *Annual Review of Linguistics* 7. 213-233.

Murray, S. (2021b). Evidentials. In *The Wiley Blackwell Companion to Semantics*, ed. by Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, & Thomas Ede Zimmermann.

N'gbala, A., & Branscombe, N. R. (1995). Mental simulation and causal attribution: When simulating an event does not affect fault assignment. *Journal of Experimental Social Psychology*, 31(2), 139-162.

Nissell, J. & Woolley, J. D. (2022) Why wearing a yellow hat is impossible: Chinese and U.S. children's possibility judgments. Presentation at the Annual Meeting of the Cognitive Development Society. Madison, WI.

Noveck, I., Ho, S., & Sera, M. (1996). Children's understanding of epistemic modals. *Journal of Child Language,* 23, 621-643.

Nyhout, A., Henke, L., & Ganea, P. A. (2019). Children's counterfactual reasoning about causally overdetermined events. *Child Development*, 90(2), 610-622.

Ozturk, O. & Papafragou, A. (2016). The acquisition of evidentiality and source monitoring. *Language Learning and Development,* 12(2), 199-230.

Papafragou, A. (1998). The acquisition of modality: Implications for theories of semantic representation. *Mind & Language*, *13*(3), 370-399.

Paul, L., Ullman, T. D., De Freitas, J., & Tenenbaum, J. (2023). Reverse-engineering the self. Unpublished Manuscript. https://doi.org/10.31234/osf.io/vzwrn

Pearl, J. (2009). *Causality*. Cambridge university press.

Perner, J., Sprung, M., & Steinkogler, B. (2004). Counterfactual conditionals and false belief: A developmental dissociation. *Cognitive Development*, *19*(2), 179-201.

Peterson, D. M., & Riggs, K. J. (1999). Adaptive modelling and mindreading. *Mind & Language*, 14(1), 80-112.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, 24(5), 586-604.

Phillips, J., & Bloom, P. (2022). Do children believe immoral events are impossible? Unpublished Manuscript.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, *114*(18), 4649-4654.

Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry*, 20(1), 30-36.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, *33*(1), 65-94.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, 145, 30-42.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*, *23*(12), 1026-1040.

Phillips, J., & Norby, A. (2021). Factive theory of mind. *Mind & Language*, 36(1), 3-26.

Portner, P. (2018). *Mood*. Oxford University Press.

Pratt, C., & Bryant, P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development*, 61(4), 973-982.

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the Logic of Causal Selection. *Psychological Review*. https://dx.doi.org/10.1037/rev0000428

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world". *Child Development*, 81(1), 376-389.

Rafetseder, E., O'Brien, C., Leahy, B., & Perner, J. (2021). Extended difficulties with counterfactuals persist in reasoning with false beliefs: Evidence for teleology-in-perspective. *Journal of Experimental Child Psychology*, 204, 105058.

Rafetseder, E., & Perner, J. (2010). Is reasoning from counterfactual antecedents evidence for counterfactual reasoning? *Thinking & Reasoning*, 16(2), 131-155.

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, 114(3), 389-404.

Redshaw, J., & Suddendorf, T. (2016). Children's and apes' preparatory responses to two mutually exclusive possibilities. *Current Biology*, *26*(13), 1758-1762.

Redshaw, J., & Suddendorf, T. (2020). Temporal junctures in the mind. *Trends in Cognitive Sciences*, *24*(1), 52-64.

Redshaw, J., Suddendorf, T., Neldner, K., Wilks, M., Tomaselli, K., Mushin, I., & Nielsen, M. (2019). Young children from three diverse cultures spontaneously and consistently prepare for alternative future possibilities. *Child Development*, *90*(1), 51-61.

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality?. *Cognitive Development*, 13(1), 73-90.

Rohwer, M., Kloo, D., & Perner, J. (2012). Escape from metaignorance: How children develop an understanding of their own lack of knowledge. *Child Development*, 83(6), 1869-1883.

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., & Apperly, I. A. (2006). Children's sensitivity to their own relative ignorance: Handling of possibilities under epistemic and physical uncertainty. *Child Development,* 77(6), 1642-1655.

Rubinstein, A. (2012). Roots of Modality. University of Massachusetts at Amherst PhD dissertation.

Rubinstein, A. (2014). On necessity and comparison. *Pacific Philosophical Quarterly*, 95, 512–554.

Rullmann, H. & Matthewson, L. (2018). Towards a theory of modal-temporal interaction. *Language* 94(2), 281-331.

Rullmann, H., Matthewson, L., & Davis, H. (2008). Modals as distributive indefinites. *Natural Language Semantics,* 16(4), 317-357.

Shtulman, A. (2009). The development of possibility judgment within and across domains. *Cognitive Development*, 24(3), 293-309.

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., & Phillips, J. (2018). Differentiating 'could' from 'should': Developmental changes in modal cognition. *Journal of Experimental Child Psychology*, *165*, 161-182.

Skinner, B. F. (1965). *Science and Human Behavior*. Simon and Schuster.

Smith, K., Battaglia, P., & Tenenbaum, J. (2023). Integrating heuristic and simulation-based reasoning in intuitive physics. Unpublished Manuscript.

Smith, K., Mei, L., Yao, S., Wu, J., Spelke, E., Tenenbaum, J., & Ullman, T. (2019). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in neural information processing systems*, 32.

Sodian, B., & Wimmer, H. (1987). Children's understanding of inference as a source of knowledge. *Child Development*, 424-433.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, 99(4), 605.

Srinivasan, G., Acierno, J. & Phillips, J. (2022). The shape of option generation in open-ended decision problems. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.*

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, 348(6230), 91-94.

Suddendorf, T., Crimston, J., & Redshaw, J. (2017). Preparatory responses to socially determined, mutually exclusive possibilities in chimpanzees and children. *Biology Letters*, 13(6), 20170170.

Tardif, T., Wellman, H. M., Fung, K. Y. F., Liu, D., & Fang, F. (2005). Preschoolers' understanding of knowing-that and knowing-how in the United States and Hong Kong. *Developmental Psychology*, 41(3), 562.

Téglás, E., & Bonatti, L. L. (2016). Infants anticipate probabilistic but not deterministic outcomes. Cognition, 157, 227-236.

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. Proceedings of the National Academy of Sciences, 104(48), 19156-19159.

Turan-Küçük, E. N., & Kibble, M. M. (2023). Three-year-olds' ability to plan for mutually exclusive future possibilities is limited primarily by their representations of possible plans, not possible events. Unpublished Manuscript. Boston University.

van Dooren, A., Dieuleveut, A., Cournane, Aílish, & Hacquard, V. (2022). Figuring Out Root and Epistemic Uses of Modals: The Role of the Input. *Journal of Semantics*, 39, 581-616.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science,* 38(4), 599-637.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. Journal of Experimental Psychology: General, 121(2), 222.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3), 655-684.

Wells, Gordon. (1979). Learning and Using the Auxiliary Verb in English. In V. Lee (ed.), *Cognitive Development: Language and Thinking from Birth to Adolescence*. London (Croom Helm), 250–70.

Wells, Gordon. (1985). *Language Development in the Pre-School Years.* Cambridge University Press. Cambridge.

Williamson, T. (2016). Knowing by imagining. In A. Kind & P. Kung (ed.), *Knowledge through Imagination*. Oxford (Oxford University Press), 113-123.

de Wit, T. C., & van Lier, R. J. (2002). Global visual completion of quasi-regular shapes. *Perception*, *31*(8), 969-984.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, 100(2), 283-301.

Wurmbrand, S. (1999). Modal verbs must be raising verbs. *Proceedings of WCCFL,* 18, 599-612.

Yalcin, S. (2016). Modalities of Normality. In N. Charlow & M. Chrisman (ed.), *Deontic Modality*. Oxford (Oxford University Press), 230-55.

Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, 119(2), 166-178.

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023). Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*.