COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2017) 1–33

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12509

Knowing When Help Is Needed: A Developing Sense of Causal Complexity

Jonathan F. Kominsky,[a] Anna P. Zamm,[b] Frank C. Keil[c]

[a]*Department of Psychology, Harvard University*
[b]*Department of Psychology, McGill University*
[c]*Department of Psychology, Yale University*

Received 24 August 2015; received in revised form 21 April 2017; accepted 2 May 2017

Abstract

Research on the division of cognitive labor has found that adults and children as young as age 5 are able to find appropriate experts for different causal systems. However, little work has explored how children and adults decide when to seek out expert knowledge in the first place. We propose that children and adults rely (in part) on "mechanism metadata," information about mechanism information. We argue that mechanism metadata is relatively consistent across individuals exposed to similar amounts of mechanism information, and it is applicable to a wide range of causal systems. In three experiments, we show that adults and children as young as 5 years of age have a consistent sense of the causal complexity of different causal systems, and that this sense of complexity is related to decisions about when to seek expert knowledge, but over development there is a shift in focus from procedural information to internal mechanism information.

Keywords: Causal mechanisms; Explanation; Deference; Cognitive Development

1. Introduction

Would you need help from an expert to repair the engine of a 747? Not to underestimate the capabilities of our readers, but we expect that for most of you the immediate and intuitive answer is "yes." Most of you will likely reach this conclusion while knowing very little about how airplane engines work, the structure of their internal mechanism, their materials or construction, etc., and yet you know something about them that tells you that you could not simply figure it out on your own. Contrast this with the same intuition about a clock. You might have an equally limited idea of the details of the internal

Correspondence should be sent to Jonathan F. Kominsky, Department of Psychology, Harvard University, William James Hall #1154, 33 Kirkland St., Cambridge, MA 02138. E-mail: jkominsky@g.harvard.edu

mechanism of a clock, but at the same time, your likely intuition is that you would need much less assistance to repair one.

These intuitions present an intriguing puzzle. On one hand, they seem to indicate some understanding of the causal systems in question. On the other hand, that knowledge seems to be sorely lacking in concrete details. Ask yourself, for example, how an electric-powered analog clock differs from a pendulum-powered one, or what components are shared between a clock and a watch. Relying entirely on your own knowledge, these questions are probably very challenging to answer. Even so, you have a clear sense that you would need less help to fix a clock than a jet engine. What exactly is it that you know, then?

Several studies have investigated how children and adults select which experts to defer to, and how these experts are evaluated. However, little work has explored the question of what prompts us to look for help in the first place. Nonetheless, in reviewing research on *how* children and adults defer to relevant experts, there are hints as to the kind of knowledge that is used to determine *when* to defer as well.

1.1. The division of cognitive labor

Children and adults rely extensively on networks of distributed knowledge (Mills, Legare, Grant, & Landrum, 2011; Sparrow, Liu, & Wegner, 2011; Wilkenfeld, Plunkett, & Lombrozo, 2016), and they demonstrate an early-developing ability to identify which experts are relevant to the task at hand. Previous work has found that, by age 5, children have a basic understanding of how broad domains of knowledge are clustered in the minds of other people (Danovitch & Keil, 2004; Keil, Stein, Webb, Billings, & Rozenblit, 2008). Even preschoolers will judge that a car mechanic is more likely than a doctor to know about how to build a tree house (Lutz & Keil, 2002). For more specific objects, younger children can infer who will know how to fix an object, versus what a novel object is called, on the basis of the kinds of knowledge each expert has demonstrated previously (Kushnir, Vredenburgh, & Schneider, 2013).

In these previous studies, the primary goal was to determine whether children were sensitive to information about informants. Children use a diverse array of information about informants when deciding whom to defer to, ranging from relevant factors like past accuracy (Koenig & Harris, 2005) to irrelevant features like niceness (Landrum, Mills, & Johnston, 2013). However, this body of work also implicitly shows that children know something about the causal systems for which they are seeking expert knowledge. For example, in order to understand that a mechanic will know more than a doctor about how to build a tree house (Lutz & Keil, 2002), young children must know something about the knowledge of the mechanic and the doctor, but also something about the process of building a tree house. Children may not know exactly what is required to build a treehouse, or how to construct one, but they must know something that allows them to say that a mechanic's body of knowledge is more relevant to the task than a doctor's. As they grow older, children draw on more elaborated expectations about the kinds of processes that are responsible for entities and mechanisms in broad domains (Keil et al., 2008).

In addition to knowing *whom* to defer to, children (and adults) must have some sense of *when* expert knowledge is needed in the first place. One obvious way to determine the need for help is to experience failure. Indeed, children will defer to an expert when they have attempted something and failed (Vredenburgh & Kushnir, 2015). However, to determine whether you needed help to fix a jet engine, you (presumably) did not go find a broken jet engine and attempt to repair it. Thus, there must be ways to recognize when deference to an expert is appropriate based on more limited knowledge and experience. What, then, does one need to know?

One way to assess the need for deference is to learn more about the experts who typically interact with a given system or entity and to examine the qualifications involved. For example, rather than knowing anything about jet engines, one might know that the people who typically repair jet engines undergo extensive training and certification, and from that information infer that one could not figure things out on one's own. With reliable indicators of training and needed qualifications, it may be relatively easy to infer one's distance from those levels of expertise.

However, knowledge about training and qualifications of particular experts may be difficult to acquire and generalize, especially for young children. When, for example, is a child likely to encounter information about training and certification of jet engine technicians? In addition, to be able to use such information for a wide range of devices and causal systems, one needs to know who is specifically responsible for interacting with each of those devices and systems. Generalizations cannot be made without also knowing something about the causal patterns involved. One could, for example, extrapolate from the expertise required for one device, such as refrigerators, to make inferences about another, such as air conditioners, but even this inference requires being able to recognize that the two devices share some causal mechanistic properties, and therefore requires some information about the causal systems in addition to the experts. Therefore, to understand how people determine when they will need help, we should determine what they know about these causal systems.

1.2. Causal mechanism knowledge (and the lack thereof)

While causal knowledge has been studied in many forms, little work has been done on the type(s) of knowledge that could support decisions about when help is needed. Humans are certainly adept at inferring causal relationships from an early age. Children are able to predict outcomes of novel causal events as early as 8 months of age (Sobel & Kirkham, 2006) and less than a year after that can make successful causal interventions (Gopnik, Sobel, Schulz, & Glymour, 2001; Gopnik et al., 2004), even for causal relationships defined by abstract relational properties (Walker & Gopnik, 2014). However, causal learning from statistical information would not necessarily support intuitions about whether (and which) expert knowledge is required, at least not on its own. In order to solve the problems of deference described above, we need some knowledge of the *means by which* a cause brings about an effect, that is, its mechanism (Ahn, Kalish, Medin, & Gelman, 1995). Merely knowing that the causal relationship exists does not entail any

knowledge of its internal mechanism. Indeed, previous work has made it clear that children do not require mechanism information to infer that a causal relationship exists (e.g., Schulz, Gopnik, & Glymour, 2007).

Even if children do not need information about mechanism to infer causal relationships, children nonetheless ask for mechanism information about things they encounter starting around 3 years of age (Callanan & Oakes, 1992) or even earlier (Hood & Bloom, 1979), and often persist in their inquiries until they receive a mechanism-oriented response (Chouinard, 2007; Frazier, Gelman, & Wellman, 2009). This preference for mechanism-based explanations continues into adulthood (e.g., Ahn et al., 1995; Johnson & Ahn, 2015). If we knew the mechanism underlying the function of a device or biological system, it would be relatively straightforward to determine if we would need help understanding it. Therefore, if the early-developing interest in mechanism translated into a deep understanding of the underlying causal systems, that understanding could be a valuable basis for intuitions about when help is required.

Unfortunately, while laypeople often believe they have such detailed causal understandings, they usually do not. This phenomenon is known as the Illusion of Explanatory Depth (IOED) (Alter, Oppenheimer, & Zemla, 2010; Rozenblit & Keil, 2002; Sloman & Fernbach, in press). Major gaps in adult causal knowledge occur not just in recall but also in recognition. For example, many adults fail to recognize the difference between a schematic of a functional bicycle and one that is completely inoperable (Lawson, 2006). Children show such illusions of understanding to an even greater degree than adults (Mills & Keil, 2004).

Despite these limitations, some kind of information about mechanism seems to persist: Mechanism information influences causal reasoning (Ahn et al., 1995; Schlottmann, 1999), and mechanism may constrain Bayesian causal learning by reshaping priors about what causal links exist or how strong they are (Griffiths & Tenenbaum, 2005, 2009). Thus, some aspects of mechanism information are preserved, but they are neither detailed nor complete (e.g., DiSessa, Gillespie, & Esterly, 2004; Straatemeier, van der Maas, & Jansen, 2008; Vosniadou, 2002). So, if most individuals do not retain deep, integrated mechanistic understandings of "the way things work" (e.g., Macaulay, 1988), what kind of information about mechanism persists, and how does it support intuitions of when to seek out expert knowledge?

1.3. Mechanism metadata

Even if one does not have a detailed understanding of the mechanism that enables a car to move, one may know, for example, that the mechanism involves metal and plastic rather than organic parts, that it involves the transfer of mechanical force, and that acceleration is not instantaneous. In other words, even if one does not know the details of how the mechanism works, one may still know roughly how much "stuff" is in the mechanism and have some ideas about the nature of that "stuff."

We can describe this type of knowledge as "mechanism metadata." Metadata simply means information about information. The term is most commonly used in the domain of

web design; in addition to the content of the page which is visible to the user, most webpages also contain "metadata" in their code which are invisible to the user but available to search engines, to allow those pages to show up in searches for terms that may be relevant but that are not actually contained in the content of the page itself (Zhang & Dimitroff, 2005). Webpage metadata is a convenient analogy for our purposes, as we propose that mechanism metadata supports the search for information in the minds of others.

This metadata may take many forms. For example, simply knowing what ontological category a causal system belongs to (e.g., artifact or natural kind) is a form of mechanism metadata, as it is information about the mechanism that requires few or no concrete details of the mechanism. More sophisticated examples might include ideas about rough causal structures (e.g., that a car is a common-cause structure, centered around the engine). The best candidate form of metadata for determining the need to defer may be summary information, that is, a sense of approximately how much "stuff" is in a causal mechanism and how diverse that "stuff" is.

Although mechanism metadata does not require detailed mechanism knowledge, it is strongly compatible with fragmentary mechanism knowledge. For example, for the common-cause structure described above, one must know about the existence of an engine, and the parts to which an engine ultimately connects (e.g., the wheels), even if one has sparse ideas concerning the intervening components and their causal interactions.

Notably, mechanism metadata should be distinguished from *just* fragmentary mechanism knowledge, or more unspecified "stuff we know about things." Rather, mechanism metadata is made up of specific types of information that share specific properties. In this initial proposal we do not attempt to construct a comprehensive taxonomy of different types of mechanism metadata, nor will we attempt to describe in any detail the cognitive processes that give rise to mechanism metadata. Rather, our goal is to specify two key properties that define certain information as mechanism metadata and focus on the observable features those properties entail.

First, mechanism metadata should be relatively consistent across individuals with similar amounts of exposure to a given causal system. If mechanism metadata is extracted from the search for information about a causal system (its mechanism or its behavior), then assuming that the information that different people are exposed to is reasonably accurate and at a similar level of detail, the retained mechanism metadata should be similar even if the (largely forgotten) details were different. For example, one person could receive information about the steering system in a car and another about the brakes, and either way much (but not all) of their resulting metadata knowledge would be the same (device, transfer of force from driver to mechanism, etc.). Thus, for populations that have been exposed to similar amounts of information at similar levels of detail, we should expect considerable consistency in the mechanism metadata they possess.

More broadly, this point highlights that mechanism metadata is not just a collection of fragmentary details, but rather some form of abstraction from those details. Indeed, when people demonstrate holes in their mechanistic knowledge, there is some variability in what they get right and what they seem to be missing (e.g., Lawson, 2006, Fig. 3 and Table 1). So, if metadata were simply a collection of people's fragmentary knowledge, it

would show little consistency across individuals. Finding consistency therefore provides evidence for metadata as a separate kind of information. However, different populations that receive different amounts or levels of detail in their mechanism information (for example, children versus adults, or experts versus laypeople) might end up with different metadata.

Second, we should expect people to have mechanism metadata for virtually every causal system they encounter. For example, while the metadata for artifacts and natural kinds would likely differ in content (e.g., kinds of material substrates or causality), the metadata should not differ in amount or type. This much follows from our definition of mechanism metadata. The types of metadata we suggest should apply equally to artifacts and natural kinds, indeed to almost all causal systems. Most causal systems belong to an ontological category, have some overall density of structure, some number of components, some diversity of components, etc.

This constraint distinguishes our proposal from an obvious alternative: If we do not make decisions about deference on the basis of domain-general metadata, then we must use domain-specific or idiosyncratic information. For example, one could propose that people use "price" to guide deference to experts when dealing with artifacts. However, information like "price" would not help guide deference for most biological systems. The key feature of our proposal is that there is something domain-general that supports deference in all domains, and it is that which we intend to test.

This constraint also distinguishes mechanism metadata from being simply "stuff we know about things" by specifying that consistency alone is not sufficient. There are some types of information about some causal systems where you might find great consistency in lay judgments (e.g., price, longevity), but that do not apply across every causal system. As such, consistent domain-specific judgments provide an excellent alternative hypothesis to mechanism metadata, but are importantly different.

Assessment of metadata intuitions is distinct from assessing information that people retain about detailed mechanism knowledge. In some respects, partial assessments along these lines already exist. For example, there is evidence for ontological category metadata by age 5, in that 5-year-olds show some ability to match biological insides to animals and device-like insides to machines (Gelman & Wellman, 1991; Gottfried & Gelman, 2005; Simons & Keil, 1995). Notably, to make these identifications, they matched images of biological systems and mechanical systems to different causal systems, even though the mechanisms portrayed were, in their details, totally inaccurate to the system. Indeed, even 8-month-old infants seem to possess rudimentary metadata of this sort, in that they expect objects that are self-propelled and behave like living things to have insides, that is, to not be hollow (Setoh, Wu, Baillargeon, & Gelman, 2013). Thus, in searching for the types of mechanism metadata that help us decide when it is necessary to find expert help, one straightforward and developmentally viable method is to probe the metadata as directly as possible, while showing that it is related to decisions about when to defer. Here, we focus on metadata related to complexity intuitions.

1.4. Causal complexity

One type of mechanism metadata stands out as an especially strong candidate for enabling children and adults to determine when they will need help: the degree of causal complexity of a mechanism. For our purposes, we will define causal complexity as an intuitive, approximate sense of the number and diversity of components in the mechanism of a causal system. If a causal system is judged to be more causally complex, it is more likely that one will need to defer to an expert to interact with it successfully.

Recent work has suggested that children and adults represent this kind of complexity, and relate it to the observed behavior of a causal system. Children as young as 4 will match images of more complex internal components (more and different components) with objects that have a greater number and/or diversity of behaviors (Ahl & Keil, 2016; Erb, Buchanan, & Sobel, 2013). It is important to note the renderings of "complexity" in these studies had no relation to the actual mechanism involved, thus ruling out detailed mechanism knowledge as an explanation for this performance. Indeed, this sense of causal complexity may not always be an accurate sense of the real number and diversity of parts within a given causal system. Even if this form of mechanism metadata is based on exposure to mechanism information, laypeople may not be exposed to enough information to accurately extract a sense of causal complexity, or they might fail to recognize non-obvious sources of causal complexity or oversimplify inferred mechanisms based on naïve causal theories (e.g., Grotzer & Tutwiler, 2014; Hmelo-Silver, 2004; Jacobson, 2001; see also http://xkcd.com/1425/).

Nonetheless, we should expect this sense of causal complexity to be applied broadly to a wide range of causal systems, and we should also expect it to be relatively consistent, at least within a given population. It is less clear whether it will be consistent across populations, or at what age it starts to be consistent. Very young children, who may be exposed to simplified mechanism metadata in idiosyncratic ways, might not extract mechanism metadata that is as consistent as school-age children or adults. Indeed, 4-year-olds may not even recognize complexity in the same way that older children do (Ahl & Keil, 2016).

Between childhood and adulthood more broadly, there may be some major shifts in what is seen as complex, as the underlying causal knowledge expands and becomes more sophisticated. Even in adults, expertise can reveal hidden complexity behind seemingly simple causal systems (e.g., Shtulman & Valcarcel, 2012). For example, a refrigerator seems simple (if opaque) until you understand the principle of the conservation of energy, at which point it becomes clear it cannot simply be generating "cold" (a concept that becomes yet more challenging when you understand what temperature actually represents). Given the potential for developmental changes in mechanism metadata knowledge, it is therefore useful to examine the development of this sense of causal complexity, especially in relation to decisions about when help is needed, particularly because such decisions may be multifaceted.

Thus far, we have only discussed decisions about when one would need help to fix something, but children are rarely required to fix things on their own. However, children

more often require help simply using many causal systems that adults can use without supervision. Seeking help for fixing may therefore be linked to different intuitions than seeking help for using. One may not see any need for help in using a refrigerator, but a need for help in fixing one. Conversely, one may need help to use chopsticks, but no help to fix them. Between these reasons and changes in exposure to mechanism information over development, there may be some developmental shifts in how complexity intuitions are grounded and linked to judgments about when help is needed.

However, the relationship between complexity and judgments of when help is needed could, in certain cases, be bidirectional. In particular, when laypeople have information about the degree of expertise needed to fix or use a mechanism but little information about the mechanism itself (e.g., a nuclear reactor), then the sense of how much "stuff" is in the mechanism could be influenced by this deference information. That is not to say that this deference information is itself a form of metadata. Rather, metadata as we have proposed it takes input from whatever information happens to be observable for a given mechanism, which in some cases may include need for deference, and metadata can then be used to "fill in" unobserved information. This is not only true of deference, the same bidirectional relationship could exist for other inputs to mechanism metadata. For example, previous work on complexity used observable behavior to guide inferences of internal complexity (Ahl & Keil, 2016; Erb et al., 2013), but in cases where something is known about the internal mechanism but nothing about the behavior (e.g., until recently, the Antikythera mechanism), more complex internal mechanism might suggest more complex observable behaviors.

1.5. The current experiments

Here, we present initial evidence for the existence of causal complexity metadata that is closely related to judgments of when to seek help. We set out with three goals for these experiments: first, to establish whether children and adults make consistent judgments of complexity in the domain of artifact devices, or if such judgments are idiosyncratic; second, to show that these judgments of complexity are related to judgments about when to seek expert help; and third, to show that this sense of complexity fulfills all of the criteria we set forward for mechanism metadata, being consistent between individuals, largely impervious to context, and applicable across a broad range of causal systems. Across all of these goals, we examine the development of this sense of causal complexity over middle childhood and into adulthood, to determine if there are notable or consistent shifts in what is seen as complex, how it relates to deference, or how consistent complexity judgments are at different ages.

Experiment 1 directly asks whether children ages 7–10 and adults possess this sense of causal complexity, and whether it is related to judgments of when help is needed. Previous work has found that children's understanding of both causal systems and the division of cognitive labor changes substantially in this age range (e.g., Danovitch & Keil, 2004; Keil et al., 2008; Mills & Keil, 2004). Thus, we might expect that we could capture similar developmental effects in a form of knowledge that is connected to both understanding

causal systems and deference to expert sources. Experiment 1 is designed to test, through measures of scale reliability, one of our criteria for this sense of causal complexity as a form of mechanism metadata, that is, testing whether it is consistent across individuals with similar amounts of knowledge. In addition, we examine whether these judgments of complexity correlate with judgments about when help is needed to interact with a causal system, either to interact with its internal mechanism, or simply to use it.

Experiment 2 further explores the nature of this sense of causal complexity, and whether it fits our definition of mechanism metadata, by testing whether adults apply it across very different causal systems. As a form of mechanism metadata, it should be applicable to any causal system, not just artificial devices. So we extend the complexity task from Experiment 1 to a new domain (human body parts). Above and beyond testing whether this mechanism metadata can be applied to these different domains of causal systems, we also test the stability of this sense of complexity, by presenting the items in these different domains as either a mixed list or two separate lists, to determine whether these judgments are robust across different contexts, or if they are made ad hoc to the particular set of items at hand. This immunity to context effects is important—if complexity judgments are subject to context in this way it would suggest that those complexity judgments are capturing domain-specific rather than domain-general information, even if we find consistency within a domain.

Experiment 3 expands on Experiment 2 by testing whether children also have a consistent sense of the complexity across different domains, particularly domains in which certain types of deference are irrelevant (i.e., you cannot ask for help using body parts). Furthermore, we examine how early this sense of complexity shows the consistency we find in adults. Using a subset of the items from Experiment 2, we elicited complexity judgments from 7- to 10-year-olds and also 5–6-year-olds, to see whether younger children, who may have more limited and idiosyncratic exposure to mechanism information, have as consistent a sense of the complexity of these items as older children and adults.

2. Experiment 1

Experiment 1 had three goals: first, to demonstrate the existence of a sense of causal complexity that fits our description of mechanism metadata, by showing that children and adults make consistent judgments of the complexity of different causal systems; second, to determine whether this sense of complexity is in fact related to decisions about when help is needed; and third, to determine *how* this sense of causal complexity relates to decisions about when to defer, and whether this relationship changes between children and adults.

We took the 16 most commonly mentioned devices from several datasets in the CHILDES utterance database and asked adults and children ages 7–10 to rate the devices' causal complexity, how much help they would need to fix the objects if they were broken, or how much help they would need to use the objects for their intended purpose. We consider "help needed to fix" as a measure of the need for deference for internal

mechanism information, and "help needed to use" as the need for deference for what we might call procedural information. To test whether complexity is indeed tied to deference, we simply correlated complexity ratings with these two deference scales.

Because this initial exploration used correlational measures, we cannot firmly establish a *causal* relationship between complexity judgments and decisions of when help is needed. In Experiment 1, our goal was to demonstrate the existence of complexity metadata that is consistent across individuals and to demonstrate that it is related to judgments of when deference is needed, while leaving deeper investigations of the direction of influence for future work.

2.1. Methods

2.1.1. Participants

We recruited 60 adults from Amazon Mechanical Turk. Adult participants received modest monetary compensation for a roughly 10-min experiment. For our child age groups, we recruited 79 7–8-year-olds (34 male, 46 female) and 62 9–10-year-olds (29 male, 33 female) from elementary schools in neighboring towns as well as a local children's museum. All child participants were rewarded with a certificate of appreciation and a small toy.

2.1.2. Stimuli

Stimuli were created from the CHILDES database (MacWhinney, 2000), specifically the Cartarette, Warren, and Evans datasets. These datasets were chosen because they represent the utterances of children in or slightly younger than the age groups examined in this study, they were within all from 20 years of when the experiments were conducted, and they all have morphological tags. The Cartarette dataset consisted of elicited speech from children ages 6, 8, 10, and adults (utterances by adults were not analyzed), discussing their everyday lives. The Warren dataset was built from recordings naturalistic interactions between a parent and child (ages 2–6, but only ages 4–6 were used in this analysis) at home. The Evans dataset was built from recordings of naturalistic interactions between pairs of 6–8-year-old children at school. This broad spread allowed for a very general sense of which objects were frequently discussed by children. Using the CLAN analysis software, lists of nouns and frequency counts were extracted. We selected the 16 most frequently occurring nouns that referred to devices with electronic and mechanical components. Stimuli for this experiment can be found in Table 1. In addition, there were four filler items that were not included in the final analysis. These four items (boomerang, gyroscope, broom, and bean-bag chair) were an exploratory set of stimuli for a future project and were not anticipated to have any impact on the ratings of the other items (see also Experiment 2).

2.1.3. Procedure

Participants were randomly assigned to one of three conditions: Complexity, help needed to fix the object if it were broken (henceforth FIX), and help needed to use the

Table 1
Stimuli generated from CHILDES data

Devices	Body parts
Airplane*	Arm*
Camera*	Blood
Car*	Elbow
Clock	Eye*
Flashlight*	Finger*
Microphone	Hair*
Microscope	Heart
Radio	Knee
Scooter	Nose*
Stereo	Shoulder
Submarine*	Skeleton*
Telephone*	Throat*
Television*	Thumb
Truck*	Toe
Vacuum cleaner	Tongue
X-ray machine	Teeth*

Note. Only device items were used in Experiment 1. * = items seen by children in Experiment 3 (children saw all device items in Experiment 1).

object for its intended function (henceforth USE). Participants in the Complexity condition were asked to rate "How complicated is this object in terms of how it works?" on a 0–100 scale going from "very simple" to "extremely complex," represented as a slider that always started at 0. In the FIX condition, participants were asked, "How much would you need help with in fixing this object, if it didn't work at all?" on a 0–100 scale from "I could do it all on my own" to "I couldn't do any of it on my own." Child participants also had the midpoint labeled "I could do about half of it on my own." In the USE condition, adult participants were asked, "How much would you need help with using this object for its primary function?" and child participants were simply asked, "How much would you need help with using this object?" using the same rating scale as the FIX condition. Notably, the wording of the FIX and USE scales left little room for interpretation. For a child, "fixing" a television could be as little as changing the channel, but specifying "fixing it if it didn't work at all" pushed the question much more toward the internal mechanisms of the object.

Adults and children in the FIX and USE conditions completed four training items to familiarize themselves with the scale before answering the question for the 20 test items. For adults, the training items asked how much help they would need to cook a hardboiled egg, how much help they would need to cook a cheese soufflé, how much help they would need to put a band-aid on a cut, and how much help they would need to conduct quadruple bypass surgery. For children, the training items were how much help they would need to make a jelly sandwich, how much help they would need to make fried chicken, how much help they would need to draw a flower, and how much help they would need to paint the *Mona Lisa*.

In addition, prior to these training items, children in all three conditions received a brief training on how to use the slider, involving an illustration of a jar and a scale that went from "Empty" at one end to "Full" at the other. There were three of these training items, and the jars displayed were always somewhere in between completely full and completely empty, to train children to use all of the scale rather than just the endpoints.

For adult participants, all test items were presented one at a time in random order in an online Qualtrics survey (Qualtrics, 2005). For child participants, items were presented one at a time in random order using a very similar Qualtrics survey presented on an iPad, using the iPad's built-in web browser. Notably, for children, the scale had no visible numerical values, as we felt that children would find this distracting. Children's responses were recorded by where they placed the slider on the scale, which could be done by dragging the slider with their fingers or tapping the point on the scale where they wanted the slider to go. For child participants, the experimenter read the question aloud for each item (e.g., "How much help would you need with fixing an airplane, if it didn't work at all?").

2.2. Results

Children were grouped into 7–8-year-olds and 9–10-year-olds for analysis.[1] We omitted child participants who failed to rate the "full" jar higher than the "empty" jar on the initial scale training items. Seven 7–8-year-olds were excluded from analyses, three for failing this exclusion criterion and four due to a software error. Three 9–10-year-olds were excluded, one for previously participating in a pilot version of the experiment and two for failing the exclusion criterion. We planned to apply a further exclusion criterion to all age groups to remove anyone who could not use the scale, operationalized as having a standard deviation of less than 10 (on a 100-point scale), which would indicate that they gave the same response (or very nearly the same response) to every single item. No participants were excluded based on this criterion in this experiment.

The final Ns were as follows: Complexity, 24 7–8-year-olds, 20 9–10-year-olds, and 22 adults; FIX, 25 7–8-year-olds, 20 9–10-year-olds, and 20 adults; USE, 24 7–8-year-olds, 19 9–10-year-olds, and 18 adults. The means for each item and scale are presented in Fig. 1a–c.

2.2.1. Scale reliability

To test whether this sense of complexity is consistent across individuals with similar levels of knowledge, we used measures of scale reliability. Scale reliability was calculated as Cronbach's α. Generally, a scale with Cronbach's $\alpha \geq 0.7$ is considered to be internally reliable (Cronbach, 1951). However, because Cronbach's α can be inflated by a large number of items (Cortina, 1993), a second measure of reliability was used to validate these results: the average single-measure intraclass correlation coefficient (ICC). The ICC is often used as a measure of inter-rater reliability with continuous scales (Bartko, 1966), and it can be tested in an F test against a null hypothesis value of 0, though the exact strength of ICCs can be difficult to interpret (Weir, 2005).

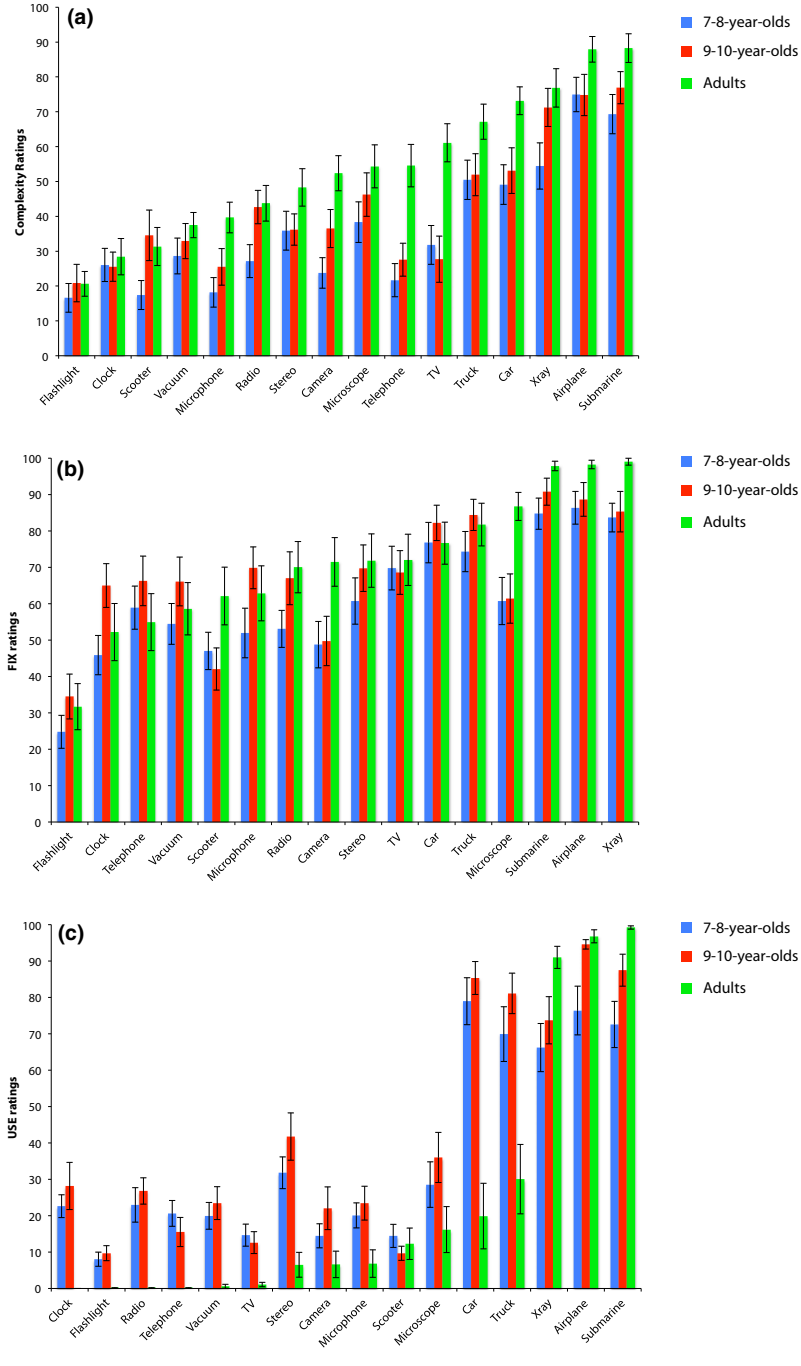

Fig. 1. (a–c) Average complexity (a), FIX (b), and USE (c) ratings for each item in Experiment 1, by age group. Note that each graph is arranged in order of adults' average rating, but this order is not the same for each measure. Error bars represent ± 1 *SEM*.

The alphas and ICCs for all three scales in each age group can be found in Table 2. Both children and adults showed good reliability for all three scales. All alphas were above the reliability threshold of 0.7 for every age group (range 0.706–0.924), and all ICCs were significant and of low to moderate strength (range 0.131–0.545, all ps < .001).

2.2.2. Correlations with complexity

Ratings were averaged for each item by condition and age group. Each age group was examined separately. First, we examined whether there was an overall correlation between these three scales. For each age group, there was a significant positive correlation between complexity and FIX (7–8-year-olds: r = .891, 9–10-year-olds: r = .750, Adults: r = .890), and complexity and USE (7–8s: r = .917, 9–10s: r = .897, Adults: r = .808), as well as a significant correlation between FIX and USE (7–8s: r = .858, 9–10s: r = .843, Adults: r = .801), all ps < .001. In short, we can safely say that judgments of complexity are closely related to decisions about when to defer, and also that these two types of deference are closely related to each other.

Given that all three scales were strongly correlated with each other, and given that we have proposed that there may be bidirectional influence between deference and complexity metadata, we sought to determine whether the relationship between complexity and deference was more related to one or the other of our deference scales. To examine this, we conducted partial correlations of complexity judgments with each deference scale while controlling for the other.

For adults, complexity was strongly and significantly correlated with FIX when USE was controlled for, r = .689, p = .005, but the correlation between complexity and USE when FIX was controlled for was non-significant r = .345, p = .207. This indicates that, for adults, FIX is more central to adults' understanding of causal complexity, either as a source of input to mechanism metadata or as a judgment informed by complexity metadata.

For children, this was not so. For 7–8-year-olds, the relationship between complexity and FIX was marginal when USE was controlled for, r = .507, p = .054, and the relationship between complexity and USE was significant when FIX was controlled for, r = .655, p = .008. This indicates that both deference scales are related to 7–8-year-olds understanding of causal complexity. For 9–10-year-olds the relationship between complexity and FIX was not significant when USE was controlled for, r = −.027, p = .923, but the relationship between complexity and USE was still significant when FIX was controlled

Table 2
Cronbach α and single-measure ICCs (in parentheses) from Experiment 1

	Complexity	FIX	USE
7–8-year-olds	0.841 (0.249**)	0.895 (0.346**)	0.812 (0.213**)
9–10-year-olds	0.843 (0.251**)	0.916 (0.406**)	0.735 (0.148**)
Adults	0.924 (0.431**)	0.950 (0.545**)	0.706 (0.131**)

**p < .01.

for, $r = .745$, $p = .001$. This is an inverse of the pattern found with adults, suggesting that for 9–10-year-old children, USE is the more central form of deference in relation to causal complexity.

However, examining each age group alone does not give us insight into whether these results represent a significant shift in the relationship between complexity and different forms of deference between age groups. It is possible to convert correlation coefficients into Z scores by using Fisher transformations, and from those Z scores, conduct a z-test comparing them (e.g., Yu & Dunn, 1982).

We examined, post hoc, whether the partial correlations above differed between age groups, resulting in six comparisons (all three possible pairs of age groups for each partial correlation). This revealed only one significant difference: The partial correlation between complexity and FIX, controlling for USE, was significantly greater for adults than 9–10-year-olds, $p = .026$. However, there was no significant difference between these groups in the partial correlation between complexity and USE when FIX is controlled for ($p = .126$), and no differences between the 7–8-year-old group and any other group (all ps $\geq .135$).

In short, there seems to be a change in the relationship between complexity and deference between older children and adults, suggesting that USE is central to the relationship between complexity and other forms of deference in 9–10-year-olds, but not in adults. Future studies with a greater number of items (and therefore greater power) might uncover further significant differences between age groups.

2.2.3. Developmental differences in complexity judgments

One key question concerns how the sense of complexity changes over development. The above analyses suggest a shift in the relationship between complexity and deference, but do not give insight into changes in the sense of complexity in itself. There are two key questions that we wished to address: First, are some age groups more internally consistent than others? That is, do adults have a more consistent sense of complexity than 7- or 9-year-olds? Second, independent of the first question, do children and adults rate different things as being complex?

With regard to the first question, Cronbach's alphas can be compared given the alpha, number of raters, and number of items, by computing a confidence interval for each alpha and conducting a chi-square test (Feldt, Woodruff, & Salih, 1987). This technique is implemented in R's cocron package (available in R or as a web-based interface; Diedenhofen & Musch, 2016). For judgments of complexity, this analysis revealed no significant between-group differences in the magnitude of alpha, $\chi^2(2) = 3.47$, $p = .18$. The 95% confidence intervals of alpha for each age group are provided in Fig. 2.

The second question is more challenging to address. Simple comparisons of means could use items as a factor with 16 levels and, crossed with three age groups, have almost as many free parameters as participants, which would violate many assumptions of a standard ANOVA and create an unreasonable number of multiple comparisons. Yet collapsing items together is likely to disguise item effects, and we do expect there to be differences between items. The most straightforward solution is to fit a linear mixed effect model to

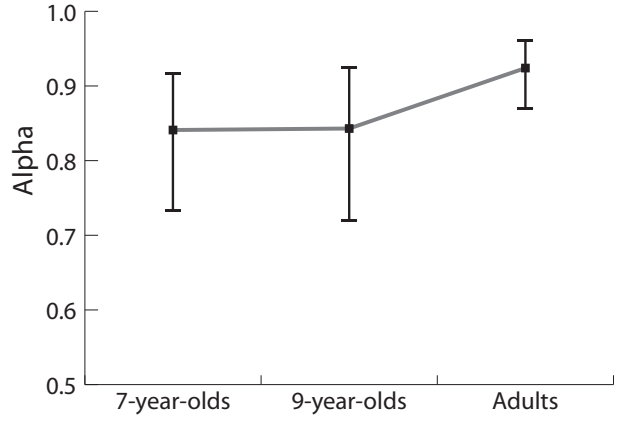

Fig. 2. Cronbach's alphas for complexity ratings in each age group in Experiment 1. Error bars represent 95% CIs.

complexity ratings that includes subject as a random factor and treats age group and item as fixed factors, with each individual rating as an observation. This allows one to determine if there is an interaction between item and age group, which would indicate that different ages rated different items as more or less complex.

Using R's afex package (Singmann et al., 2016), we fitted complexity ratings to a linear mixed effect model with age group and item as fixed factors and subject as a random factor, and then conducted an F-test on the fixed effects and interactions to determine if there was a significant interaction between age and item.[2] This analysis identified no significant main effect of age group, $F(2, 422.7) = 4.19$, $p = .13$, a significant main effect of item, $F(15, 945) = 53.33$, $p < .001$, and critically, a significant interaction between item and age group, $F(30, 945) = 1.87$, $p = .003$. This interaction indicates that different age groups considered different items to be complex.

To explore this interaction further, we computed least-square means contrasts between age groups for each item using R's lsmeans package (Lenth, 2017). In essence, this allows us to conduct pairwise comparisons between age groups for each item independently, but it does so on the basis of model predictions that are corrected for the influence of subject and main effects. These comparisons used Tukey-corrected p-values for 48 contrasts. The full set of comparisons can be found in the supplementary online materials. To summarize, adults significantly differed from 7- to 8-year-olds on seven items: Telephone, TV, Camera, Car, X-ray, Microphone, and Submarine, all ps $\leq .02$. Adults significantly differed from 9-10-year-olds on three items: TV, Telephone, and Car, all ps $\leq .02$. Notably, 7–8-year-olds did not differ significantly from 9- to 10-year-olds on any items, all ps $> .05$. In other words, adults differed from children in what objects they regarded as more or less complex, but within the age range of 7–10, children did not differ from each other.

2.3. Discussion

Experiment 1 showed that children ages 7–10 and adults make highly consistent judgments of causal complexity, at least within their own age groups. On its own, this is a striking result. Previous work has consistently found that adults have little or no detailed understanding of how causal systems like these actually work (e.g., Lawson, 2006; Rozenblit & Keil, 2002), and children even less so (Mills & Keil, 2004). Even if some of our participants had more extensive knowledge of some of these causal systems, it is implausible that they all would. As such, these results provide initial evidence that these complexity judgments reflect a form of mechanism metadata, by demonstrating high consistency across individuals.

In addition, Experiment 1 demonstrated that complexity is indeed related to decisions about when to defer to experts, though we did not establish a direction of influence. However, we did find preliminary evidence for an intriguing shift between middle childhood and adulthood in how complexity relates to deference: The relationship between complexity and FIX was mediated by USE in older children, but not in adults. This suggests that the relationship between different forms of deference and complexity data changes between ages 9–10 and adults, and that USE may play a more central role for children. Intuitively, this makes some sense: Children often need help using devices, but they are rarely called upon to fix them. Alternatively, such USE information may be more readily available to children than internal mechanism information, and it would therefore be more likely to inform complexity metadata. This opens opportunities for future work exploring the direction of influence and whether USE is in fact a more relevant form of deference for children than FIX.

Finally, Experiment 1 provided evidence that what is seen as complex may shift between children and adults. The results revealed disagreements between children and adults for a handful of items (notably TV, Telephone, and Car), but give no information as to why these particular items should differ between children and adults. More systematic work will be needed to determine whether these differences are reliable and the underlying principles behind them. However, it is important to first establish that these complexity judgments do in fact reflect our definition of mechanism metadata.

3. Experiment 2

Experiment 1 showed that a sense of causal complexity is related to decisions about when help is needed, but this sense of complexity itself requires further exploration. In particular, we proposed that this sense of complexity is a form of mechanism metadata and defined a set of properties that we expect it to possess. Experiment 1 showed that it has at least one of these properties (consistency among individuals exposed to similar amounts and types of mechanism information). A second critical property concerns its applicability to a wide range of causal systems. Notably, this is something that should distinguish complexity from FIX and USE judgments. USE would be difficult to apply to

many biological causal systems, as would FIX in some cases. Therefore, in Experiment 2, we once again asked participants to judge the causal complexity of different causal systems, but we added a set of items from a very different domain: human body parts. We predicted that people should be just as consistent in their complexity ratings for these items as they were for the items used in Experiment 1.

Furthermore, one could argue that the consistency found in Experiment 1 does not reflect any kind of general metadata but instead a series of ad hoc comparisons across the particular items we used. We are definitely not suggesting that this metadata knowledge is stored in the form of a 100-point scale, but at the same time, it should be more general and stable than a series of ad hoc comparisons. We expect that people have a sense of how complex a causal system is even when no other systems are presented to compare it to, and that it is the same sense of complexity independent of what other items are available for comparison. If it were only comparisons between specific causal systems, we should expect these ratings to be prone to powerful context effects; that is, they would change if items were presented together in different combinations. Therefore, in addition to testing whether these complexity ratings can be applied to different causal systems, in Experiment 2 we also tested whether presenting the new body part items and the device items used in Experiment 1 in mixed lists or separate lists would change the consistency of the complexity judgments or in fact change the judgments themselves.

3.1. Methods

3.1.1. Participants

We recruited 42 workers from Amazon Mechanical Turk who did not participate in Experiment 1, for modest monetary compensation. The goal was to get 20 for each condition, with the two extra participants recruited due to experimenter error.

3.1.2. Stimuli and procedure

Using the same methods and dataset as Experiment 1, we extracted the 16 most commonly mentioned body parts from the CHILDES database. In addition, we used the same 16 device items from Experiment 1.

Participants were randomly assigned to one of two context conditions, “separate” and “mixed.” In both conditions, participants were given the same instructions as adults in Experiment 1’s complexity condition. However, rather than the items being presented one at a time, they were presented in groups of 16, so each participant only saw two pages worth of items. In the separate condition, one page had the 16 device items, and the other had the 16 body part items, so that items in each domain were only seen on the same page with other items in the same domain. In the mixed condition, we randomly selected 8 items from each domain (using random.org) to make a list of 16 items, with the other list simply being the other 8 items from each domain. Thus, in the mixed condition, each item was on a page with items from its own domain and the other domain. The two “mixed” lists can be found in Table 3. Within each page, the items were presented in random order, and the order of pages was randomized between participants.

Table 3
Stimuli in "mixed" list condition from Experiment 2

Page A	Page B
Arm	Finger
Eye	Elbow
Heart	Hair
Nose	Knee
Teeth	Shoulder
Throat	Skeleton
Thumb	Tongue
Toe	Blood
Airplane	Car
Camera	Clock
Flashlight	Microphone
Radio	Microscope
Stereo	Scooter
Submarine	TV
Telephone	Truck
Vacuum cleaner	X-ray machine

Note. The order of pages and the order of items within each page were randomized.

3.2. Results

We excluded any participants who failed to use the scale, operationalized as a standard deviation of less than 10 in either domain (regardless of condition). This excluded a total of 8 participants, 5 for having a standard deviation of less than 10 in the body parts domain, 2 for having a standard deviation of less than 10 in the device domain, and 1 for having a standard deviation of less than 10 in both domains. These results therefore report data from 34 participants in total, 18 in the separate condition and 16 in the mixed condition. Means for each item in each condition are presented in Fig. 3.

3.2.1. Context effects

To examine whether context impacted ratings, we once again fit a linear mixed-effect model to complexity ratings, with list condition and item as fixed factors and subject as a random factor. This analysis found a significant main effect of item, $F(32, 992) = 11.96$, $p < .001$, no significant effect of condition ($\beta_1 = -2.65$, $SE = 8.65$), $F(1, 121.69) = 0.09$, $p = .76$, and no significant interaction between item and condition, $F(31, 992) = 0.84$, $p = .72$. However, a null effect is not necessarily evidence in favor of the null hypothesis, and in this case it would be useful to have stronger evidence that the condition X item interaction is not a significant predictor of rating. It is possible to generate a Bayes factor for a linear mixed effect model using R's BayesFactor package and the generalTestBF function (Morey, Rouder, & Jamil, 2015). The BF_{01} for a given term in the model is the posterior likelihood that the data were generated by a model in which that term was not a significant predictor. This analysis revealed the Bayes factor for the null hypothesis to be $BF_{01} = 4120.95$, indicating that the null hypothesis (that the condition X item interaction

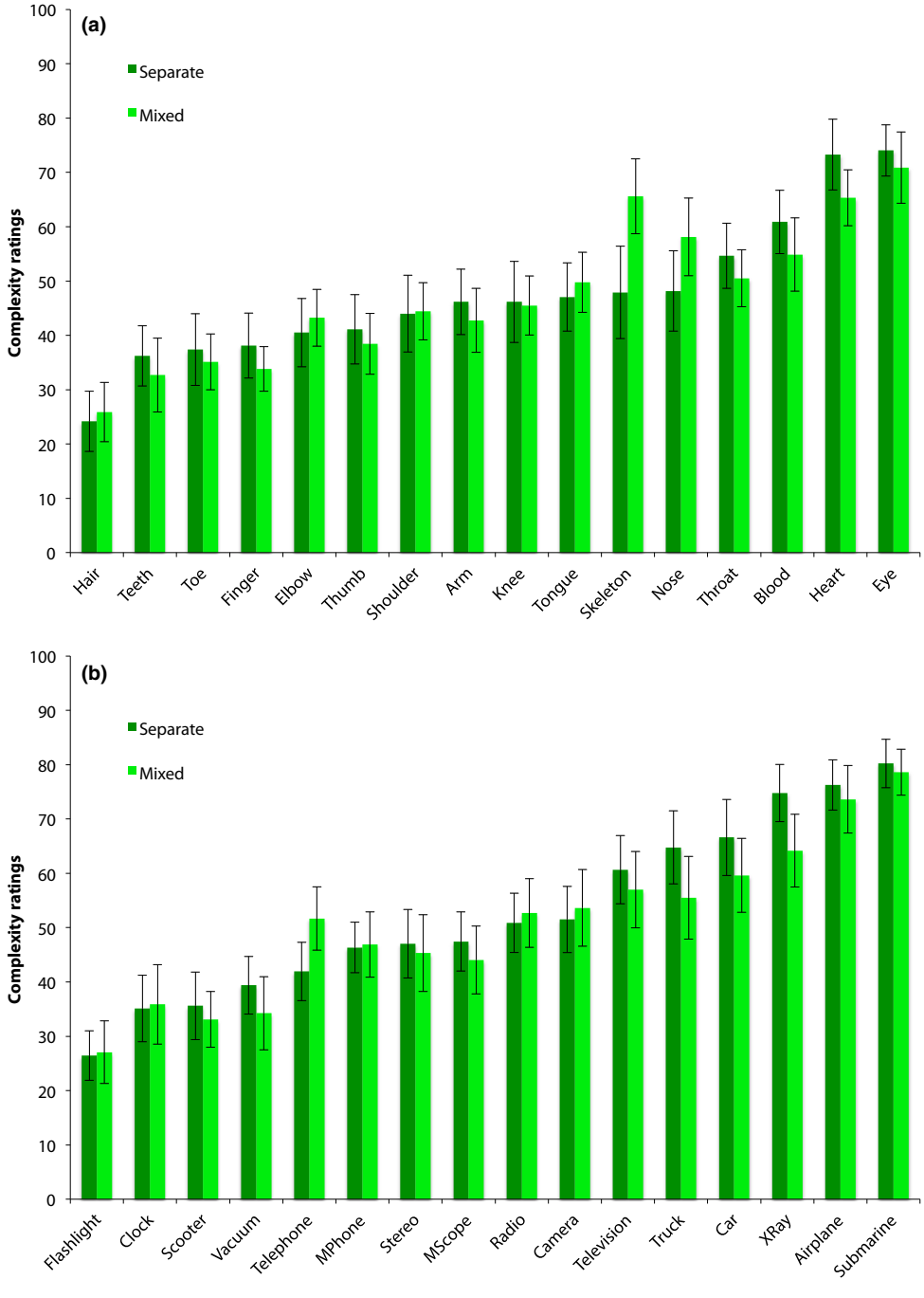

Fig. 3. (a and b) Complexity ratings from Experiment 2 for body parts (a) and devices (b), each in order of mean complexity rating in the Mixed condition. Error bars represent ± 1 *SEM*.

had no effect on ratings) is over 4,000 times more likely than the alternative (that the interaction did have an effect on ratings).

3.2.2. Scale reliability

Cronbach's α and single-measure ICCs are presented in Table 4, examined separately by condition and domain, but also combined across both dimensions. Reliability was extremely high in every possible subset, with all alphas greater than 0.9 and all ICCs significantly greater than 0 at $p < .001$. In short, participants in this experiment were highly consistent in their ratings of the complexity of these items, regardless of the domain of the item or the context in which the item was presented.

3.3. Discussion

This sense of causal complexity fits our account of mechanism metadata well. It applies just as easily to items in the domain of body parts as it does to artifacts, and the context in which the items are presented has no detectable impact on judgments of their causal complexity.

This expansion into the domain of biological systems does introduce complications into the findings of Experiment 1. In particular, it is difficult to say how the types of deference we investigated might map on to the domain of biology. One could maybe "fix" body parts, in a sense, but it is far more difficult to find an analogous question to how one might "use" body parts. On the basis of the correlations reported in Experiment 1 (with no clear evidence of directionality), this raises the possibility that children might have a different understanding of the complexity of body parts, or no understanding at all. However, if this sense of causal complexity is truly a form of mechanism metadata, then it should not matter that it is not applicable to one particular form of deference. Rather, children should just as easily have access to metadata about biological causal systems, though it is still possible that their sense of the causal complexity of biological systems differs from that of adults.

4. Experiment 3

Experiment 3 asked whether children's understanding of causal complexity extends to biological causal systems. In addition, Experiment 1 found that even 7-year-olds show a

Table 4
Cronbach α and single-measure ICCs (in parentheses) from Experiment 2

	All Items	Body Parts	Devices
Separate lists	0.976 (0.556**)	0.971 (0.673**)	0.953 (0.561**)
Mixed lists	0.956 (0.405**)	0.910 (0.386**)	0.941 (0.498**)
Combined	0.968 (0.483**)	0.952 (0.552**)	0.946 (0.524**)

**$p < .01$.

consistent sense of causal complexity, but earlier work has demonstrated both that the search for mechanism knowledge starts much earlier than 7 years of age (e.g., Hood & Bloom, 1979), and that younger children are adept at navigating the division of cognitive labor in a way that implies sensitivity to some forms of mechanism metadata (e.g., Lutz & Keil, 2002). Therefore, we might expect this sense of causal complexity to be present at earlier ages as well. So, in this experiment, in addition to seeing whether older children's complexity judgments extended to the domain of human body parts, we examined whether a sense of complexity is present in 5–6-year-old children and if it is different from that of older children or adults.

4.1. Methods

4.1.1. Participants

We recruited 24 adults from Amazon Mechanical Turk who had not participated in previous experiments. Adult participants received modest monetary compensation for a roughly 10-min experiment. For our child age groups, we recruited 29 5–6-year-old (18 male, 11 female), 25 7–8-year-olds (34 male, 46 female), and 22 9–10-year-olds (13 male, 9 female) from elementary schools in neighboring towns as well as two local children's museums. All child participants were rewarded with a certificate of appreciation and a small toy.

4.1.2. Stimuli

For adults, the items were the same as in Experiment 2. For children, we took the 8 items from each domain that appeared most frequently in CHILDES. These are the items starred in Table 1.

4.1.3. Procedure

The procedure was identical to that of the complexity rating condition of Experiment 1, except for the new stimuli.

4.2. Results

We applied the same exclusion criteria as in previous experiments, and excluded any participants with a standard deviation of less than 10 in either domain. This excluded four adults, nine 5–6-year-olds, five 7–8-year-olds, and two 9–10-year-olds. We replaced excluded participants until we had 20 in each age group. For the adults, we examined both the full set of 32 items used in Experiment 2, and the subset of 16 items that were presented to children in this experiment. For reliability, we report values for both sets of items. For comparisons between age groups, we only examine the 16 items seen by all age groups in this experiment. Results can be found in Fig. 4.

4.2.1. Scale reliability

Reliability results are presented in Table 5. As in Experiment 1, 7–10-year-old children and adults showed a great deal of consistency with others their own age, though 9-10-year-

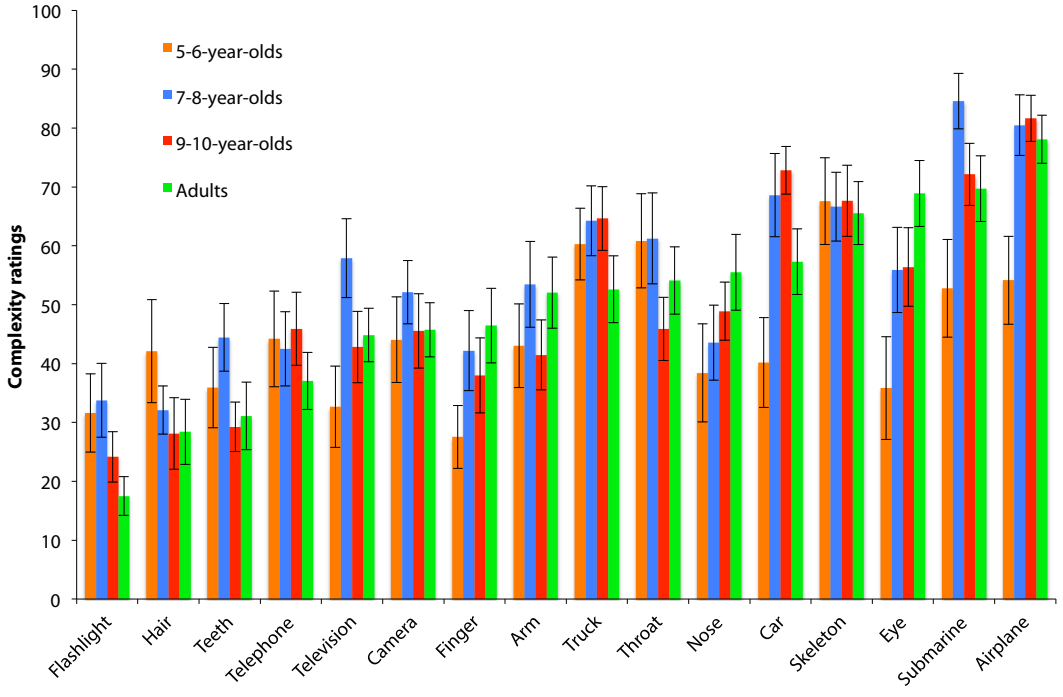

Fig. 4. Average complexity ratings by item for each age group in Experiment 3. Only the 16 items seen by all age groups are presented here, though adults saw the same 32 items that were used in Experiment 2. Error bars represent ± 1 *SEM*.

olds slightly less so (but still above the 0.7 threshold with highly significant ICCs). However, while 5–6-year-olds had an alpha just above 0.7 when items from both domains were examined, they showed subthreshold consistency when each domain was examined separately. Given that alpha as a measure is affected by the number of items, alphas for smaller numbers of items from the same scale will typically be lower (though not always, see the 9–10-year-olds). Nonetheless, 5–6-year-olds showed significant (but small) single-measure ICCs. Overall, this pattern suggests that younger children do have a sense of causal complexity, but perhaps less robustly than older children or adults (but see below).

4.2.2. Developmental differences

As in Experiment 1, there are two key questions. First, whether there is a difference in the internal consistency of some age groups relative to others. Second, whether there are differences in what is seen as complex.

Using the same analysis as Experiment 1, we conducted a chi-square test of the effect of age group on the alphas for the full 16-item scale. This analysis revealed a significant effect of age, $\chi^2(3) = 8.03$, $p = .045$. The alphas and 95% CIs are presented in Fig. 5. Post hoc pairwise comparisons found that 5–6-year-olds were significantly less consistent than 7–8-year-olds, $\chi^2(1) = 5.26$, $p = .022$, and marginally less consistent than adults,

Table 5
Cronbach α and single-measure ICCs (in parentheses) from Experiment 3

	All Items	Body Parts	Devices
5–6-year-olds	0.726 (0.142**)	0.675 (0.206**)	0.528 (0.123**)
7–8-year-olds	0.914 (0.398**)	0.869 (0.452**)	0.841 (0.398**)
9–10-year-olds	0.746 (0.155**)	0.764 (0.288**)	0.776 (0.303**)
Adults (kid items)	0.883 (0.319**)	0.903 (0.538**)	0.854 (0.422**)
Adults (all items)	0.932 (0.300**)	0.959 (0.592**)	0.901 (0.362**)

** $p < .01$.

$\chi^2(1) = 2.92$, $p = .088$, suggesting that the 5–6-year-olds' sense of complexity was less robust than those of (some) older children and adults. Surprisingly, 9–10-year-olds were also significantly less consistent than 7–8-year-olds in this experiment, $\chi^2(1) = 4.63$, $p = .03$. No other comparisons were significant (all ps $> .1$). The high consistency of 7–8-year-olds is similar to Experiment 1, but the lower consistency of 9–10-year-olds is unexpected. However, none of these effects are very strong, and Fig. 5 makes clear that the 95% CIs are quite broad. Future work examining differences in consistency between age groups may benefit from greater power, both in number of participants and number of items.

To examine differences in what items were seen as complex, we fit a linear mixed effect model to the complexity ratings with age group and item as fixed factors and subject as a random factor. This analysis found a significant main effect of age group, $F(1, 646.28) = 4.41$, $p = .004$, a significant main effect of item, $F(15, 1140) = 9.10$, $p < .001$, and a significant interaction, $F(45, 1140) = 2.17$, $p < .001$.

As in Experiment 1, we then conducted least-square means pairwise contrasts of the effect of age group on each item. The only significant contrasts were between 5- and 6-year-olds and older age groups: 5–6-year-olds differed from 7-8-year-olds in their ratings of Submarine, Car, Airplane, and TV; from 9-10-year-olds in their ratings of Car and Airplane;

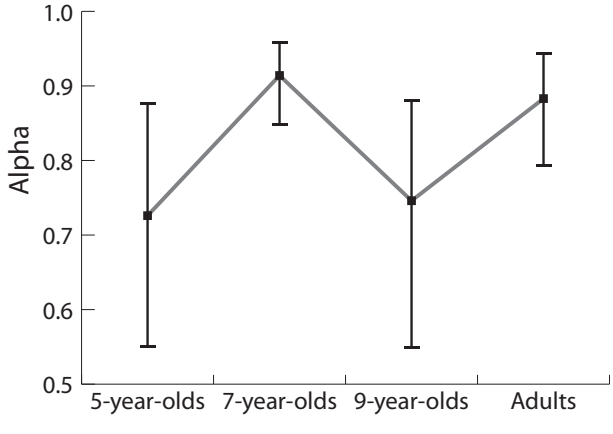

Fig. 5. Cronbach's alphas for each age group in Experiment 3. Error bars represent 95% CIs.

and from adults in their ratings of Eye and Airplane, all ps $\leq .033$. Older children and adults did not differ significantly from each other on any items, all ps $> .2$.

It is surprising that there were no differences between older age groups, given the results of Experiment 1. There are two major differences between this experiment and Experiment 1: the inclusion of 5–6-year-olds, and the items themselves (replacing eight device items with body part items). We therefore conducted two post-hoc analyses to attempt to determine whether these changes to the experiment could explain the difference in the results.

Perhaps including 5–6-year-olds somehow masked the age differences between the groups we examined in Experiment 1, if nothing else by increasing the number of pairwise comparisons from 48 to 64, thereby making the corrected p-values that much more conservative. This is quite trivial to test, we simply fit another linear mixed-effect model and excluded the 5–6-year-old group, thereby performing the exact same comparison that we performed in Experiment 1, but with new items and new participants. This analysis revealed no effect of age group, $F(2, 388.67) = .10$, $p = .91$, a main effect of item, $F(15, 855) = 11.60$, $p < .001$, and only a marginally significant interaction, $F(30, 855) = 1.46$, $p = .054$. Pairwise comparisons with this more restricted sample found no significant contrasts, all ps $\geq .11$. Therefore, merely including the 5–6-year-olds does not explain the discrepancy with Experiment 1.

To verify that changing the items did not also change the ratings, we compared ratings of the eight items that appeared in both Experiment 1 and Experiment 3 across experiment, in the three age groups that were in both experiments, using a linear mixed-effect model with experiment and item as fixed factors and subject as a random factor. In all three age groups, there was a main effect of item, ps $< .001$, no main effect of experiment, ps $> .14$, and critically, no interaction, ps $> .14$. In other words, there was no indication that ratings of the device items changed in any age group as a result of the different set of items, which aligns with the results of Experiment 2.

These analyses suggest that this is simply a non-replication of the developmental effects we found in Experiment 1 due to type-II error (or Experiment 1's results were due to type-I error, though by Experiment 1's p-values this is by definition unlikely). However, further work will be needed to establish whether there are robust differences in complexity judgments between age groups.

4.3. Discussion

Children ages 7–10, much like adults, have a sense of complexity for biological causal systems as they do for artifact causal systems. Younger children, 5–6 years of age, also show a similarly consistent sense of causal complexity, though marginally less robust than that of older children and adults. Again, these results alone are remarkable. It would have been entirely reasonable to expect that children would have difficulty providing consistent ratings for items across such radically different domains, especially domains for which some forms of deference information may not apply. These results further indicate

that causal complexity fits our definition of mechanism metadata, applying universally across very different domains.

Children ages 5–6 differed from older children and adults in what items they regarded as being more or less complex. However, we failed to replicate the differences found in Experiment 1 for what is seen as complex among older children and adults, for reasons that are as yet unclear. It could be as simple as type-II error, or it could be something about the inclusion of body part items that changes ratings across age groups, though our analyses suggest that the latter is unlikely. In either case, the solution will be for future work to test specific hypotheses about *how* these judgments of complexity might differ between ages, and which items are or are not likely to highlight such differences. While our results leave some intriguing puzzles, our primary goal was to establish the existence of complexity as a form of mechanism metadata across middle childhood and adulthood, and for that goal Experiment 3 is successful.

5. General discussion

How do you know that you would need help to repair a jet engine, without knowing how a jet engine works? In three experiments, we suggested that children ages 7–10 (and perhaps as young as 5) and adults have a sense of causal complexity that is closely related to such decisions. We provide evidence that this sense of causal complexity is a form of mechanism metadata that is highly consistent across individuals with similar levels of exposure to mechanism information, and broadly applied to many different causal systems across different ontological domains. The current work provides just the first stage of empirical support for the existence of mechanism metadata, and many future studies are needed to work out the details. However, these studies do demonstrate that both adults and children have strong intuitions about causal complexity despite lacking the supporting mechanistic details.

The relationship between complexity and deference is intriguing, and seems to change over development. In particular, "help needed to use" seems to be related to complexity in a different way for older children and adults, in that USE seems to mediate relationship between complexity and other forms of deference for 9–10-year-old children, but not at all for adults. It is not that internal mechanism information is irrelevant to these children's sense of complexity, however. Their judgments of complexity are just as consistent for domains in which "help needed to use" is a nonsensical question. Yet, in their day-to-day lives, children may simply have little need or opportunity to fix things, but do often need help using things.

More broadly, the direction of the relationship between complexity and deference may simply be a matter of what kind of information is available for a given causal system. If children have no information about internal mechanism, but know they need help to use something, they may base complexity judgments on that information. If they know something about how a causal system works but have no personal experience trying to use it

or fix it, their sense of complexity could inform decisions about whether they would need help.

We also found preliminary evidence that this sense of causal complexity changes over the course of middle childhood and into adulthood. At 5–6 years of age, children are capable of making complexity judgments but are slightly less consistent than 7–8-year-old children and adults. If this difference in consistency proves to be robust, it could indicate more idiosyncratic exposure to mechanism (or other) information, or it could indicate that the cognitive architecture to extract mechanism metadata is not well-developed at this age. Both of these factors likely contribute, but consideration of previous work favors the idiosyncratic exposure interpretation. Children as young as 4 make consistent judgments of relative complexity in forced-choice tasks on the basis of information about the behavior of a given mechanism (Ahl & Keil, 2016; Erb et al., 2013). In those studies, children were explicitly provided with standardized information about a causal system. When they were all exposed to the same information, they were relatively consistent in their complexity judgments (and in what kinds of information they used to determine complexity).

These previous studies suggest that some of the cognitive tools for extracting mechanism metadata are present by age 5, and the developmental changes in consistency that we observe in Experiment 3 could reflect something about the information they receive and not the conclusions they draw from it. Prior to the standardized curricula of the school years, the amount and type of mechanism information children encounter may depend heavily on their parents or siblings, both in their tendency to provide mechanism information and in the kinds and quality of mechanism information provided. Thus, there may simply be huge variations in the kind of information children receive (e.g., Callanan & Oakes, 1992).

It is less clear whether there are developmental differences in which causal systems are judged to be more or less complex. Experiment 1 found a strong developmental difference in the ratings of certain items between children and adults, but the child groups did not differ from each other. However, we failed to replicate this result in Experiment 3 with the same ages (though 5–6-year-olds did differ from older children and adults). In all likelihood, if such differences do exist, they will depend heavily on the specific causal systems in question. For example, consider the ratings of the "flashlight" item, which appeared in every experiment: Every age group in every experiment provided very similar ratings, between 15 and 30 on our 100-point scale. Future work exploring developmental differences in judgments of causal complexity would probably want to avoid using "flashlight" as one of the items. On the other hand, causal systems like "car" or "eye" seem to show more variability. Ultimately, we will need a more detailed theory of how exactly this complexity metadata is constructed for different causal systems, and how the input or the processing of that input might change over development, before we can make clear predictions about when children and adults will disagree.

In addition, a further alternative exists for any of these developmental effects: There could be a large shift in the kind of mechanism metadata children and adults acquire, but due to a cohort effect rather than cognitive development. Artifacts that children currently

encounter typically are made of opaque electronics, whereas adults may have had more experience with systems that had more visible mechanical and electrical parts, or at least parts that could be more readily accessed and manipulated. While we did not collect age data from the adult participants, the average age of workers on Amazon Mechanical Turk is approximately 32 (Mason & Suri, 2012). With an age gap of 22 years between the average age of the MTurk sample and our oldest child participants, perhaps today's adults had a genuinely different experience of internal mechanism during development. Thus, the question arises as to whether adults in 10–20 years will show the same judgments as adults do now.

5.1. Mechanism metadata

While we set out to understand decisions about when to seek help, the idea of mechanism metadata can be applied much more broadly. As noted previously, existing work on the division of cognitive labor has indirectly provided evidence for the existence of a few different kinds of mechanism metadata. Here, we have done more than provide a convenient label for describing this kind of knowledge, we have made concrete proposals for what properties we should expect mechanism metadata to have. In particular, the two key features that we proposed and tested here are consistency within a population and universal applicability. More broadly, we have suggested that mechanism metadata can help resolve the seeming contradiction of the early-developing search for mechanism information, and the absence of such information even in adults.

In this project we have primarily demonstrated the existence of a particular form of mechanism metadata and laid out a set of properties that we expect to be shared by all forms of mechanism metadata. We have suggested the existence of other forms of mechanism metadata, some based on existing work (e.g., ontological category; Lutz & Keil, 2002) and some that are ripe for further exploration (e.g., approximate causal structure; Strickland, Silver, & Keil, 2017). We also provided evidence that mechanism metadata is closely related to deference, and even particular types of deference that have seldom been explored before.

Several questions remain, however. For example, does mechanism metadata guide what level of expertise you seek out? Could you ask a medical student or would you need to ask a doctor? These sorts of questions are important to our understanding of how we navigate the division of cognitive labor.

Another set of questions address our understanding of mechanism information, and information about causal systems more generally. We did not directly examine how children and adults acquire a sense of complexity in this set of experiments. Such a study would require manipulating the information children and adults are exposed to, while we were interested in what they already knew about everyday things in the world. What kinds of information are needed to extract this metadata, and what elements of that information actually inform the content of the metadata? With regard to complexity, Ahl and Keil (2016) found that young children use information about the number of functions a (fictional) device can complete to gauge its complexity (see also Erb et al., 2013), and by age 6 they will also use the *diversity* of functions as a cue to complexity (e.g., a device

that picks two kinds of flowers is simpler than one that picks flowers and berries). Notably, these studies indicate that mechanism information is not actually necessary to extract some forms of mechanism metadata, behavior is sufficient. However, as these studies were forced choice, they might not have led to the same general sense of complexity we find in our experiments. It seems likely that some categories of behavior would lead to inferences of greater or lesser complexity, but what aspects of those behaviors are relevant, or what you need to know about those behaviors, merits further investigation.

Finally, while we feel we have provided strong evidence for the existence of mechanism metadata, there are two ways one could describe mechanism metadata, and our results do not clearly distinguish between them. One view would be that this mechanism metadata is how information is represented in memory. That is, when we asked participants for complexity ratings, they accessed a stored magnitude-like representation of complexity that they already possessed for each of these causal systems and applied it to our 100-point scale. Alternatively, complexity metadata may not be represented in memory but constructed on demand from fragments of information that are. That is, when asked how complicated a causal system is, we access a set of fragmentary knowledge about that causal system and apply some generative process to it that produces our judgment of causal complexity.

Both possibilities are compatible with our definition of mechanism metadata. Either way, we have to propose some cognitive process that is highly consistent across individuals. The question is simply whether that process takes place during encoding or only when required to address a question. In fact it would be extremely difficult to distinguish these possibilities, but it is also not necessary to do so to conduct further investigations of mechanism metadata and deference. Investigating what types of information influence complexity or other forms of metadata does not require knowing whether that metadata is formed immediately on encoding or only constructed later, and we will gain insight into the underlying process either way.

6. Conclusion

Children and adults have the remarkable ability to know a great deal about a causal system while simultaneously knowing very little. They know when to find expert knowledge and where to look for it on the basis of few and fragmentary details, supported by (and perhaps supporting) a rich library of metadata knowledge. The properties of this mechanism metadata, and the cognitive processes that support it, are a rich avenue of future exploration, to better understand how we navigate a complex causal world.

Acknowledgments

This research was supported by NSF Grant DRL 1561143 to Frank C. Keil. Jonathan F. Kominsky was supported during data analysis and the writing of this manuscript by

NIH/NICHHD fellowship 1F32HD089595-01. The authors thank the school districts of Bethel, Brookfield, Meriden, and New London, CT, for their assistance with this project, as well as Stepping Stones Children's Museum, Norwalk, CT. We would also like to thank Beau Wittmer, Matthew Roth, Jessica Zhang, Katarzyna Hitczenko, Julie Merriam, and Luke Berszakiewicz for their assistance with coding and analysis; Taryn Bipat and Sinjihn Smith for assistance with data collection; and Randall Henner for assistance implementing the linear mixed-effect model analyses.

Notes

1. We divided the age ranges into 7–8 and 9–10 to provide comparable age comparisons to previous studies of children's understanding of the division of cognitive labor (e.g., Danovitch & Keil, 2004; Keil et al., 2008). In this experiment, there were no differences between these two age groups, and post hoc analyses collapsing both child age groups generated results qualitatively identical to the ones reported here.
2. Note that there is still some debate in the statistics community as to how best to determine the degrees of freedom for these *F*-tests (see https://cran.r-project.org/web/packages/lme4/lme4.pdf [retrieved 1/26/17], pg. 91). In the absence of a perfect solution, we have opted for a commonly used one that at least gives an interpretable answer. Notably this analysis treats the number of observations as subjects * items rather than number of subjects, which both licenses its use in Experiment 2 (in which there would be more free factors than number of subjects otherwise) and explains the high estimated degrees of freedom.

References

Ahl, R. E., & Keil, F. C. (2016). Diverse effects, complex causes: Children use information about machines' functional diversity to infer internal complexity. *Child Development*, *88*(3), 828–845.

Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*(3), 299–352.

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*(3), 436–451. https://doi.org/10.1037/a0020218

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, *7*(2), 213–233.

Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, *72*(1), vii-ix, 1–112; discussion 113-126. https://doi.org/10.1111/j.1540-5834.2007.00412.x

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98. https://doi.org/10.1163/156853711X591279

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Danovitch, J. H., & Keil, F. C. (2004). Should you ask a fisherman or a biologist?: Developmental shifts in ways of clustering knowledge. *Child Development*, *75*(3), 918–931.

Diedenhofen, B., & Musch, J. (2016). Cocron: A web interface and R package for the statistical comparison of cronbach's alpha coefficients. *International Journal of Internet Science*, *11*(1), 51–60.

DiSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, *28*(6), 843–900.

Erb, C. D., Buchanan, D. W., & Sobel, D. M. (2013). Children's developing understanding of the relation between variable causal efficacy and mechanistic complexity. *Cognition*, *129*(3), 494–500. https://doi.org/10.1016/j.cognition.2013.08.002

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103.

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation. *Child Development*, *80*(6), 1592–1611. https://doi.org/10.1111/j.1467-8624.2009.01356.x

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition*, *38*(3), 213–244.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and bayes nets. *Psychological Review*, *111*(1), 3–32. https://doi.org/10.1037/0033-295X.111.1.3

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), 620–629.

Gottfried, G. M., & Gelman, S. A. (2005). Developing domain-specific causal-explanatory frameworks: The role of insides and immanence. *Cognitive Development*, *20*(1), 137–158. https://doi.org/10.1016/j.cogdev.2004.07.003

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. https://doi.org/10.1016/j.cogpsych.2005.05.004

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716. https://doi.org/10.1037/a0017201

Grotzer, T. A., & Tutwiler, M. S. (2014). Simplifying causal complexity: How interactions between modes of causal induction and information availability lead to heuristic-driven reasoning. *Mind, Brain, and Education*, *8*(3), 97–114. https://doi.org/10.1111/mbe.12054

Hmelo-Silver, C. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science*, *28*(1), 127–138. https://doi.org/10.1016/s0364-0213(03)00065-x

Hood, L., & Bloom, L. (1979). What, when, and how about why: A longitudinal study of early expressions of causality. *Monographs of the Society for Research in Child Development*, *44*(6), 1–47.

Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity*, *6*(3), 41–49. https://doi.org/10.1002/cplx.1027

Johnson, S. G. B., & Ahn, W. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*, *39*, 1468–1503.

Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive Science*, *32*(2), 259–300. https://doi.org/10.1080/03640210701863339

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*(6), 1261–1277. https://doi.org/10.1111/j.1467-8624.2005.00849.x

Kushnir, T., Vredenburgh, C., & Schneider, L. A. (2013). "Who can help me fix this toy?" The distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental Psychology*, *49*(3), 446–453. https://doi.org/10.1037/a0031649

Landrum, A. R., Mills, C. M., & Johnston, A. M. (2013). When do children trust the expert? Benevolence information influences children's trust more than expertise. *Developmental Science*, *16*(4), 622–638. https://doi.org/10.1111/desc.12059

Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & Cognition*, *34*(8), 1667–1675.

Lenth, R. (2017) lsmeans: Least-squares means. R package version 2.25-5.

Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, *73*(4), 1073–1084.

Macaulay, D. (1988). *The way things work*. Boston: Houghton Mifflin. Retrieved from Library of Congress or OCLC Worldcat.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*, Vol. 2. Mahwah, NJ: Lawrence Erlbaum Associates.

Mason, W., & Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, *44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, *87*(1), 1–32.

Mills, C. M., Legare, C. H., Grant, M. G., & Landrum, A. R. (2011). Determining who to question, what to ask, and how much information to ask for: The development of inquiry in young children. *Journal of Experimental Child Psychology*, *110*(4), 539–560. https://doi.org/10.1016/j.jecp.2011.06.003

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-2.

Qualtrics. (2005). [Computer Software]. Provo, UT: Qualtrics.

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Schlottmann, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology*, *35*(1), 303–317.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322–332. https://doi.org/10.1111/j.1467-7687.2007.00587.x

Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(40), 15937–15942. https://doi.org/10.1073/pnas.1314075110

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*(2), 209–215.

Simons, D. J., & Keil, F. C. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition*, *56*(2), 129–163.

Singmann, H., Bolker, B., Westfall, J., Aust, F., Højsgaard, S., Fox, J., Lawrence, M. A., & Mertens, U. (2016). afex: Analysis of Factorial Experiments. R package version 0.16-1.

Sloman, S., & Fernbach, P. (in press). *The knowledge illusion: Why we never think alone*. New York, NY: Riverhead Books.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, *42*(6), 1103–1115. https://doi.org/10.1037/0012-1649.42.6.1103

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*(6043), 776.

Straatemeier, M., van der Maas, H. L., & Jansen, B. R. (2008). Children's knowledge of the earth: A new methodological and statistical approach. *Journal of Experimental Child Psychology*, *100*(4), 276–296. https://doi.org/10.1016/j.jecp.2008.03.004

Strickland, B., Silver, I., & Keil, F. C. (2017). The texture of causes and effects: Domain specific biases in causal reasoning. *Memory and Cognition*, *45*(3), 442–455

Vosniadou, S. (2002). Mental models in conceptual development. In L. Magnani and N. J. Nersessian (Eds.), *Model-Based reasoning* (pp. 353–368). New York, Springer. https://doi.org/10.1007/978-1-4615-0605-8_20

Vredenburgh, C., & Kushnir, T. (2015). Young children's help-seeking as active information gathering. *Cognitive Science*, *40*(3), 697–722. https://doi.org/10.1111/cogs.12245

Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science*, *25*(1), 161–169. https://doi.org/10.1177/0956797613502983

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research/National Strength & Conditioning Association*, *19*(1), 231–240. https://doi.org/10.1519/15184.1

Wilkenfeld, D. A., Plunkett, D., & Lombrozo, T. (2016). Depth and deference: When and why we attribute understanding. *Philosophical Studies*, *173*(2), 373–393.

Yu, M. C., & Dunn, O. J. (1982). Robust tests for the equality of two correlation coefficients: A Monte Carlo study. *Educational and Psychological Measurement*, *42*, 987–1004.

Zhang, J., & Dimitroff, A. (2005). The impact of metadata implementation on webpage visibility in search engine results (part II). *Information Processing & Management*, *41*(3), 691–715. https://doi.org/10.1016/j.ipm.2003.12.002

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Figure S1. Residual box plot for the model used to analyze Experiment 1, by age group.

Figure S2. Residual box plot for the model used to analyze Experiment 2, by context condition.

Figure S3. Residual box plot for the model used to analyze Experiment 3, by age group.

Table S1a–c. Pairwise comparisons of the effect of age group for each item in Experiment 1.

Table S2a–f. Pairwise comparisons of the effect of age group for each item in Experiment 3. *$p < .05$.