Computational Brain & Behavior
https://doi.org/10.1007/s42113-024-00196-7

ORIGINAL PAPER

Plausibility and Early Theory in Linguistics and Cognitive Science

Giosuè Baggio[1] · Aniello De Santo[2] · Nancy Abigail Nuñez[3,4]

Accepted: 28 December 2023

Abstract
Various notions of plausibility are used in cognitive science to argue for or against the "goodness of theories." However, plausibility remains poorly understood and difficult to analyze. We review debates in the philosophy of science on uses of plausibility in the assessment of novel scientific theories as well as recent attempts to formalize, reform, or eliminate specific notions of plausibility. Although these discussions highlight important concerns behind plausibility claims, they fail to identify viable notions of plausibility that are sufficiently different from other criteria of "good theory," such as prior probability or external coherence. We survey uses of plausibility in linguistics and cognitive science, confirming that plausibility is often a proxy for other criteria of good theory. We argue that the need remains for concepts of plausibility that can be employed to assess the quality of proposals at the early stages of theory development when other criteria are not yet applicable. We identify two such notions: one relating to formal constraints on theories and another capturing initial epistemic consensus, if not necessarily convergence on the truth, about the target system in a community of inquiry. We briefly assess the specificity and added value of these notions of plausibility relative to other criteria for good theory.

Keywords Plausibility · Early theory · Computation · Cognitive science · Linguistics

Introduction

Cognitive scientists often use notions of plausibility — computational, cognitive, neural, biological, etc. — to evaluate theories, models, or hypotheses. Plausibility considerations are seldom accompanied by definitions and metrics (in contrast with, e.g., probability), and absolute claims ("X is plausible") tend to prevail over comparative judgments ("X is more plausible than Y"). Plausibility is frequently used as a substitute or an umbrella term for wonted criteria of the "goodness of theories," like verisimilitude, empirical adequacy, or external coherence. What is plausibility, and what functions can it serve in cognitive science?

Here, we speculate that the usefulness of epistemologically autonomous notions of plausibility, as distinct from other standards of "good theory," is a function of knowledge of the target system available at a given stage of inquiry. Simplifying, the more is known about a system, the more one can rely on standard criteria to evaluate theories, models, or hypotheses, and the less profitable it is to appeal to plausibility as such. In particular, the development of theories renders concepts like *internal coherence* and *prior probability* applicable; experimental research programs make available criteria such as *verisimilitude*, *empirical adequacy*, or *posterior probability*; discoveries from adjacent research fields make notions of *external coherence* relevant. In those circumstances, plausibility considerations typically boil down to informal judgments involving exclusively or largely one or more of the standard criteria. Conversely, when little is known about the target system and when formal theories, models, and hypotheses are not yet available or are only then being set up, the standard criteria may not be applicable: there are no theoretical constructs yet to be evaluated for empirical adequacy, internal or external coherence, verisimilitude, prior or posterior probability, etc., but the need remains to assess *early stages* of theory development. Plausibility, suitably construed, may serve precisely this

✉ Giosuè Baggio
giosue.baggio@ntnu.no

[1] Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway

[2] Department of Linguistics, University of Utah, Salt Lake City, USA

[3] Universidad Panamericana, Mexico City, Mexico

[4] Institute of Philosophy, Czech Academy of Sciences, Prague, Czech Republic

Published online: 17 January 2024

function. The question then becomes how to characterize plausibility, such that it can assist us in evaluating, comparing, and selecting theories, models, or hypotheses when the standard goodness-of-theory criteria are not yet applicable. We hope to convince readers of the importance of this question, regardless of how one then proceeds to answer it.

We also hope to get the discussion started on specific concepts that could do the job. We will identify two candidate notions of plausibility: one relating to formal arguments of *theoretical invariance* and *computational tractability* of cognitive functions; another capturing *initial epistemic consensus*, if not projected future convergence on the true or final theory, about the target system by a community of inquiry. We will discuss these notions in greater detail in sections "Invariance and Tractability: Plausibility and Formal Theory" and "Community and Inquiry: Logic and Pragmatics of Plausibility." Before we get there, we will provide a tentative characterization of early theory ("Early Theory: A Partial Case-Based Typology") and then revisit some debates on and uses of plausibility in the philosophy of science ("Plausibility in the Philosophy of Science") and in linguistics and cognitive science ("Plausibility in Linguistics and Cognitive Science").

Early Theory: A Partial Case-Based Typology

Traditionally, the objects of primary interest for philosophy of science have been "our best theories" or in any case "mature theories" (Psillos, 1999). For successful, advanced theories, such as Einstein's general relativity or Darwin's evolution by natural selection, the question of plausibility does not really arise: other criteria for theory assessment are generally used, such as empirical adequacy or external coherence. As our focus shifts to the *early stages* of theory building or development, our criteria for assessing theories also shift — and then plausibility, along with a few other criteria, might become relevant. But what is "early theory"? We will not attempt a definition here. Rather, we will offer a partial typology, based on more or less specific cases to which plausibility considerations could apply.

(i) In some areas of psychological science, many theories are either *informal* or *weak* for predictive or explanatory purposes (Meehl, 1992b; van Rooij & Baggio, 2021, 2020). Psychological theories typically have a *shorter life cycle* than theories in other areas of research (Meehl, 2002): this is not a sign of rapid development, but rather of the premature demise of theories.

(ii) Fields that have been impacted by the *replication crisis* are characterized by a surplus of inconsistent empirical results and by a lack of mechanistic or formal theories that can help select, organize, and explain such results. Biomedicine is an example. The scarcity of sufficiently developed theories is consistent with a high number of false or improbable hypotheses under test, which is a natural explanation for low replication rates (Bird, 2021).

(iii) In fields that rely on formal modeling (e.g., mathematical linguistics and computational cognitive science), theory development might be held back by the fact that (a) the techniques to prove that theories are *equivalent* or *notational variants* (Johnson, 2015) are abstract and removed from the aspects of theories that other practitioners usually care about and (b) it is difficult to set formal criteria to select between theories with comparable empirical coverage, even when non-equivalence has been established.

(iv) Some *complex cognitive or social phenomena* may be easy to identify but difficult to investigate systematically. Methodological or other challenges limit the available database and the type and scope of empirical theories that can realistically be built. Language acquisition is an example: known constraints on experimental and observational studies make it difficult to build theories of the implicit knowledge that infants and children develop over time and to assess competing theories, where they exist.

(v) In mathematics, and in other fields of formal science, *conjectures* may be proposed and subsequently proved for ever larger, yet still finite samples of the relevant objects (e.g., Golbach's conjecture). In these cases, there is no empirical basis for assessing the conjecture inductively and no general theorem to replace the conjecture. However, as partial proofs or heuristic or probabilistic arguments accumulate, a conjecture may gain credibility. Conjectures are early theories in mathematics, to which goodness criteria different from strict theoremhood apply.

In all these cases, theories are or remain at early stages of development, because they are informal, weak, or too easily abandoned (i), largely absent in the face of irreproducible results or improbable claims (ii), difficult to identify or select via (non)equivalence arguments (iii), challenged by complexity and methodological constraints (iv), or stuck at the conjectural stage (v). In such circumstances, and in cases where new theories begin to emerge for novel or known phenomena (we discuss examples in sections "Plausibility in the Philosophy of Science" and "Plausibility in Linguistics and Cognitive Science"), standard criteria for the goodness of theories are less immediately applicable and plausibility considerations become relevant. Clarity is then needed on one or more notions of plausibility, distinct from other criteria and applicable to a wide range of possible early theory scenarios.

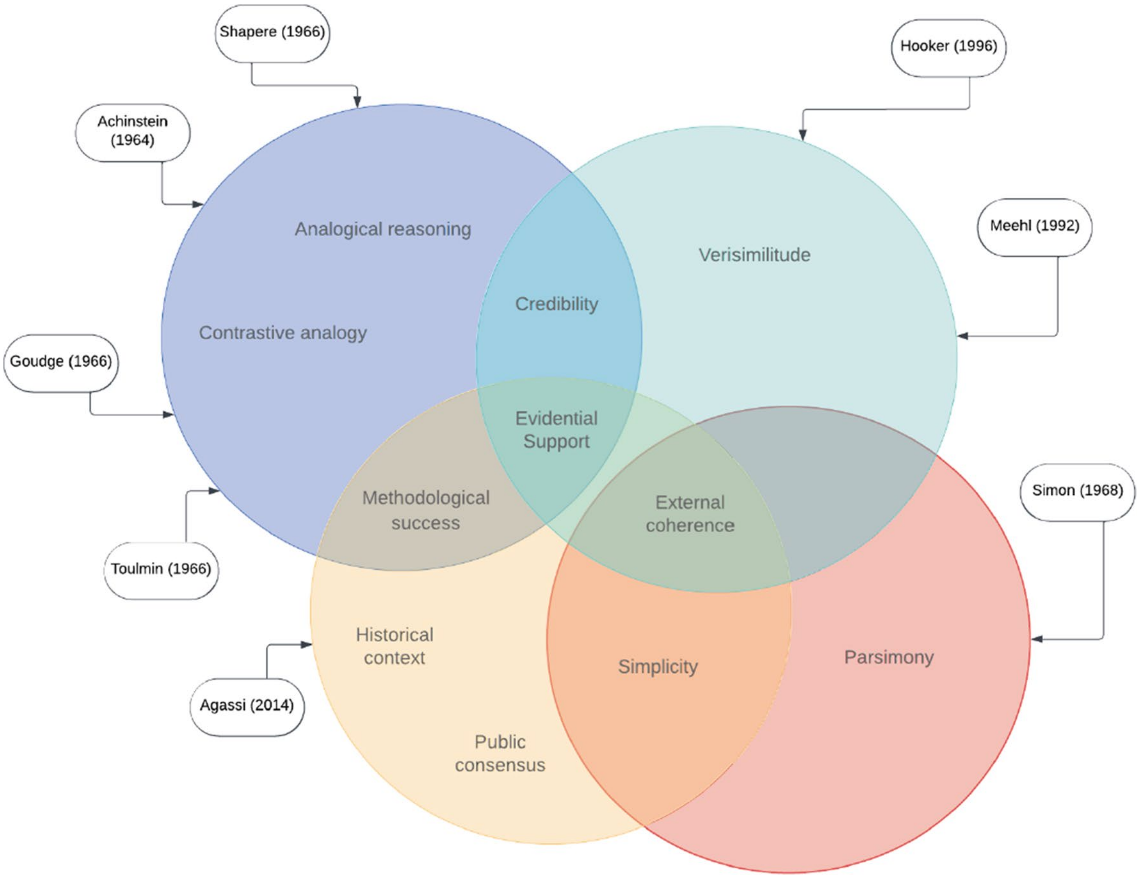

Fig. 1 A conceptual map showing the relationships between plausibility (colored circles) and other criteria of good theory according to different authors in the philosophy of science ("Plausibility in the Philosophy of Science")

Plausibility in the Philosophy of Science

Debates in the philosophy of science have shown that plausibility is entangled in a constellation of concepts and criteria of good theory and that it is occasionally equated with such concepts and criteria (Fig. 1). Achinstein (1964), and recently Bartha (2010), among others, have linked the plausibility and pursuitworthiness of early formulations of hypotheses to the credibility of *analogical relations and arguments* from established facts. Crucially, from analogical relations and from the existence of evidential support for one set of facts, nothing follows about the probability of the other set: analogy is not sufficient to *justify* a novel hypothesis or additional credible assumptions. Analogical reasoning, however, is vindicated in practice by the many instances in which it has led to new discoveries: e.g., the electromagnetic field is associated with a particle — the photon —, which "made it plausible to suggest" that the nuclear force field may also be associated with a particle — and indeed pions were later discovered. Shapere (1966) considers two examples of the uses of analogy in the formulation of novel hypotheses: Liebig's "vital force" theory, via a *contrastive analogy* with chemical forces like "cohesion and affinity" *vs* gravitation or magnetism, and Huygens' wave theory of light, by analogy with sound waves and their propagation through air. Huygens' proposal was vindicated historically to an extent that Liebig's was not, but they were both deemed plausible: the "degree of plausibility" of each proposed analogy depends on the "degree of success" of the relevant concepts and structures in the original domain (chemical forces or sound waves), "independently of any positive factual evidence in its [i.e., the analogy's] favor." In his reply to Shapere, Goudge (1966) argues that plausibility is a *methodological* idea, not an epistemological one: the initial credibility or promise of novel analogically derived hypotheses cannot be justified (as could, e.g., the prior or posterior probabilities of tested hypotheses), but reflects an assessment of viable "moves" for an investigator in context. These moves can suggest broad *classes of hypotheses* (instantiating kinds of analogical relations), but *not specific hypotheses*, which demand independent justification. Toulmin (1966) draws a different lesson from Shapere's arguments. Plausibility undercuts a hard divide between (inductive) logic and pragmatics: "reasons can be given, and judicially appraised" for taking seriously a hypothesis as plausible. Judicial appraisal involves not the use of inductive or abductive logic, but rather of "case law" reasoning, primarily precedent, to establish the applicability of the relevant concepts. The idea that credible analogy lends a hypothesis plausibility, emerging from the philosophy of science in the mid-1960s, is still at play in debates in cognitive science ("Plausibility in Linguistics and Cognitive Science"). Philosophers' attention, however, has shifted to understanding the role of analogies in supporting pursuitworthiness judgments, theoretical unification, and model transfer across domains (Nyrup, 2020). None of these is strictly specific to early theory: for example, pursuitworthiness may be assessed not only for emerging theories, but also for accepted ones, and even for rejected theories (Šešelja & Straßer, 2013; Shaw, 2022).[1]

Others have related plausibility to *verisimilitude*. This work highlights a general preoccupation with *realism*, or the capacity of (early) theories to track the truth. Here too, parallels with established facts motivate (without justifying) claims of plausibility: this is not unlike structural or conceptual analogy in Achinstein and Shapere, and builds on "relevant similarity with existing, sufficiently entrenched ontology, plus empirical adequacy and fruitfulness" (Hooker 1996, pp. 650-651). The latter two criteria clearly do not apply to early theory, and the former seems to amount to a version of *external coherence* intended to anchor new theories to accepted scientific structures. An alternative perspective is developed by Meehl. Meehl (1992a) notes that plausibility arguments tend to be weaker than proof or evidential support (but may be strengthened, if certain conditions are met): they specify "conceivable" hypotheses that should at least avoid "extreme" (very small) prior probabilities. Meehl (2002) writes that "the count of plausible theories for most fact domains is rather limited" (p. 342), though he does not list plausibility among his criteria for theory appraisal: his list comprises forms of parsimony, a theory's ability to derive a variety of facts, and two types of reducibility (see also Meehl 2004). "Initial plausibility" is part of a separate list of "additional criteria," which also contains, for example, fruitfulness, fertility, elegance, and rigor of the theoretical derivations. Meehl (2002, 2004) notes that these additional criteria are not widely accepted by either scientists or philosophers, they are difficult or impossible to quantify, they are not obviously "truth-correlated" (verisimilitude), and they may even be reducible to the main attributes. However, he does not say which of these four issues applies to "initial plausibility." For Meehl, the problem with theories in psychology is that *there are too many of them*, which have been disconfirmed, falsified, or abandoned for a variety of reasons: this he takes as an indication that a field is in a "primitive state" (Meehl 2002, p. 342). The task is to find ways of severely testing ("appraising") existing theories, and the better ones will be those with greater predictive capacity (the derivation of facts; see above). It is then clear why, in Meehl's perspective, neither initial plausibility nor any of the additional criteria play a prominent role in theory assessment.

Another view is given by Simon (1968), who relates plausibility to the *simplicity* of hypotheses that fit data patterns to a reasonable approximation. A hypothesis is plausible if it is "not inconsistent with our everyday general knowledge," if it is "already known (or strongly suspected) to be not far from the truth," and if its "subsequent empirical falsification would be rather surprising." The emphasis is here shifted from the ontological (see above) to the epistemological plane, which is where compatibility is assessed, and from what scientific theories reveal about the world (entities, structures, etc.) to "everyday general knowledge."

Agassi (2014) contrasts plausibility with proof and probability and notes that it fails to meet the standards of justification of both. Plausible or reasonable ideas may well be "preconceived views," but they are not prejudices, if one is willing to discard them for valid reasons. Agassi links plausibility to *context*, arguing that plausibility is a useful concept for making sense of how historically ideas may be believed at a given moment and rejected later, and to *community*, where what is plausible hinges on public knowledge, or is publicly accepted or permitted, or is a product of public debate. Emphasis on the community here can be contrasted with *phenomenological accounts*, where plausibility judgments are driven by a "sense or feeling of understanding," produced by relevant *explanatory hypotheses*. One problem with this account (see also Trout 2002) is that some early theories may, by definition, lack hypotheses that are sufficiently advanced to be explanatory or conducive to understanding. And if early explanatory hypotheses were available, they would only have *potential* (*vs* actual) explanatory power: plausibility would have to be based on the latter criterion, which lacks a working definition and a satisfactory analysis in the current philosophy of science.

From our discussion so far, four main propositions emerge: (1) what is assessed for plausibility are *early theories*, for which (counter-)evidence is yet to become available (Achinstein, Shapere, Meehl, Simon); (2) what gives initial plausibility to a theory is a *relation* of analogy (Achinstein, Shapere), similarity (Hooker), or consistency (Simon) with existing, established scientific or everyday knowledge; (3)

[1] Some philosophers of science may argue that there should not be any epistemic restrictions on pursuit: all proposals should be treated as equally viable forerunners of success. The concern is that "even the most seemingly trivial pursuitworthiness criterion would have inhibited some of the greatest scientific research programs in history" (Shaw 2022, 110). However, some scientific contexts, so-called "urgent science," in which there is a practical or moral reason to obtain results within a particular time frame, may demand pursuitworthiness judgments. This is another point of difference between plausibility and pursuitworthiness.

plausibility considerations cannot justify hypotheses but may *vindicate* them in practice, in those communities of inquiry that adopt them for methodological or substantive reasons (Achinstein, Shapere, Goudge, Toulmin, Agassi); and (4) plausibility is a *virtue* of (early) theories: given the choice, it is preferable (if not more rational) to consider and pursue relatively *more* plausible hypotheses.

Plausibility in Linguistics and Cognitive Science

Plausibility considerations have been used in linguistics and cognitive science to assess the relevance and applicability of certain formal frameworks to the study of the human mind or brain at different levels of analysis. We will consider three types of plausibility considerations — biological, cognitive, and computational — and discuss some of their uses in evaluating research on connectionist networks, Bayesian cognitive science, and models of tractable cognition. The aim here is to review a few illustrative uses of plausibility in these fields and to clear the space for the more positive contribution we try to make in sections "Invariance and Tractability: Plausibility and Formal Theory" and "Community and Inquiry: Logic and Pragmatics of Plausibility."

Biological Plausibility: Connectionism and Artificial Neural Networks

In the early days of connectionism, and possibly since McCulloch & Pitts (1943), the plausibility of artificial neural networks (ANNs) was predicated on a number of properties — i.e., computation by distributed units, activation thresholds, and weighted connections — that ANNs were thought to share with biological brains. Rumelhart's 1989 classic assertions that ANNs are "neurally inspired" and that computation in such systems is "brain-style computation" underscore the use of (contrastive) analogy in arguing why in cognitive science ANNs are preferable to other models of computation (e.g., the von Neumann architecture). A *relational* notion of plausibility, of the kind introduced in the "Plausibility in the Philosophy of Science" section, therefore applies here too. However, the analogy did not target the brain as such, but classical *theories of brain function* (Cichy & Kaiser 2019). Secondly, the analogy was not intended to capture all *structural properties* of biological neural networks, but only those that suffice to support the kinds of computations that brains appear to carry out (see McCulloch & Pitts' 1943 five "physical assumptions," p. 188). ANNs are thus plausible to the extent that they achieve a degree of *functional similarity* (Cichy & Kaiser 2019) with biological brains. The goal of connectionism has never been to model the brain's actual anatomy and physiology, but to reproduce and study aspects of *biological information processing* by exploiting sufficient properties of the underlying substrate. That is also why, even as knowledge in neurobiology grew (e.g., with findings on synaptic plasticity, neuron types, the functional role of inhibition, local *vs* global connectivity) and as the structural analogy between ANNs and biological brains further fell away, functional similarity claims were never disavowed (Stinson 2020). *Graceful degradation* is one example: damage to parts of an ANN generally leads to partial or minor performance losses, which "mimics the human response in many ways and is one of the reasons we find these models [...] plausible" (Rumelhart 1989, p. 231). Another example is the capacity of ANNs to uncover *latent structure* in data, which is what brains must also do and arguably what explains the recent successes of deep learning models in predicting and exploring cortical activity (e.g., Ramakrishnan et al. 2015).

That being said, the functional analogy between ANNs and brains breaks down for key processes such as learning. Biological and machine learning are different in terms of initial conditions, input requirements, learning trajectories, learning outcomes, and robustness of learned representations (but see Achille et al. 2019 for one intriguing exception). Backpropagation is a major reason why ANNs are implausible models of biological learning: weight changes propagate backwards through the entire network (this has "no plausible physical interpretation" and violates "basic properties of locality"; Grossberg 1987, p. 50), and over long time periods, backpropagated gradients tend to vanish or explode (Bengio et al. 1994; LeCun et al. 2015). Here, different responses are possible, ranging from arguing that backpropagation is not as implausible as it may appear, to developing more plausible learning rules (e.g., Dror & Gallogly 1999; Richards et al. 2019; Yang & Wang 2020). These responses tend to assume that plausibility is, again, an analogy or similarity with relevant functional properties of biological brains. But analogy is not the same as empirical adequacy. Backpropagation lacks a neurobiological counterpart, a corresponding process in the brain. Analogy is supposed to allow for a more permissive assessment than empirical adequacy of the compatibility of backpropagation with our best theories of biological brains and with the facts they are based on. For aspects of an early theory or model that lack an empirical counterpart, as indeed backpropagation, the question arises as to which criteria one can use: plausibility as analogy or similarity has been one of them.

In spite of the field's adoption of a standard concept of plausibility, not everyone would agree that biological plausibility is a *virtue* of theoretical constructs. For example, Mewhort (1990) concedes that theories or models in psychology need "tuning" at the physiological level. But he also writes that "biological plausibility must start with behavioral accuracy" (p. 161): anatomical and physiological facts about the brain do not compel us to accept or reject computational architectures (e.g., favoring connectionism over von Neumann's), and such decisions must be based on other criteria of the goodness of theories, such as a theory's capacity to explain behavior. In a similar spirit, Dror and Gallogly (1999) argue that biological plausibility is largely irrelevant for analyses at Marr's (1982) computational and algorithmic levels: analyses that contradict or disregard biological facts can still be useful in characterizing a computational problem. More recently, Love (2021) has expressed concern about asking biological plausibility questions about levels of analysis other than the implementational level or about mechanisms that are not reducible to biological components or interactions. His proposal emphasizes *coherence* and *continuity* between levels of analysis or mechanistic explanation: models at different levels must be assessed for their capacity to explain variance in data and by how well they satisfy mutual constraints.

All these perspectives presuppose that *sufficient empirical data*, beyond generic insights or observations, are available, or that candidate *explanatory theories* or *models* exist. On this assumption, we agree that standard criteria for evaluating theories should be used, instead of biological plausibility. As noted in the "Early Theory: A Partial Case-Based Typology" section, however, there exist numerous cases where data are abundant, but insufficiently structured to support theory development or selection, or where theories are not yet at a stage where they can make discriminating use of data, e.g., in prediction. In such cases, a relational or analogical concept of plausibility may prove useful. McCulloch and Pitts (1943) relied on generic observations about biological brains, which sufficed to justify claims of analogy or functional similarity for ANNs, but not to defensibly present ANNs as *theories of brain function* with predictive and explanatory power vis-à-vis empirical observations from anatomy or physiology. We do not concur with Love (2021) that "the term biological plausibility should be dropped": it should be used when other criteria are not applicable, in the type of cases discussed, and it should be applied at appropriate levels of analysis.

Cognitive Plausibility: Bayesianism and Beyond

Moving to the computational and algorithmic levels ("Computational Plausibility: Tractable Cognition"), we find again notions of cognitive plausibility with limited applicability to early theory. These amount to a theory's or a model's ability to match or fit (1) human input/output behavior or (2) human errors, or (3) to take into account human-like constraints (e.g., limited rationality) (Kennedy's 2009): (1) and (2) quite clearly presuppose ongoing empirical research programs and sufficiently developed theories so that the outcomes of measurements can be compared to theory or model predictions. Consider again ANNs. The lack of structural similarity between ANNs and the brain implies that these models are not anchored to a single level of analysis (e.g., the implementational level). Questions of *cognitive* plausibility then arise, separate from questions of biological and neural plausibility. The cognitive plausibility of ANNs is taken to depend on their capacity to "match" observed human behaviors or neural activity (Oota et al. 2022; Michaelov et al. 2021; Branco et al. 2020). This type of work is concerned with *correlations* between measures drawn from ANNs and observed variables in brain and behavior, but it does not consider the plausibility of the processes or representations induced by ANNs, of the training methods or samples (e.g., compatibility with features of realistic input to human learners), or of the specific mechanisms or resources (i.e., context, memory) implemented in a model. Thus, architectures that make assumptions compatible with popular theories of human cognition (e.g., limited contextual information, bounded memory) are occasionally discarded in favor of models showing higher correlation with human behavior or brain activations on the relevant tasks (e.g., Merkx & Frank 2021; Michaelov et al. 2021). Plausibility therefore collapses into empirical accuracy which, as mentioned, is rarely applicable to early theory.

Other authors adopt a broader notion of cognitive plausibility, arguably closer to verisimilitude and realism. In a discussion of computational models of language acquisition, Phillips and Pearl (2015) situate requirements of cognitive plausibility in contrast to the idealizations or approximations models have to make and tie it essentially to empirical validity: a model is more plausible to the extent that it approximates the *acquisition task*. This is defined not just in terms of successful output or predictions, but also in terms of the *units of representation* and *types of constraints* incorporated into a model. Plausibility partly guarantees realism: plausible models are *empirically testable* and have *greater explanatory power* compared to other models. Plausibility is argued to help foster these additional theoretical virtues, instead of the latter being requirements on plausible models. This perspective too presupposes some knowledge of the target system, to which the model is compared: from this comparison follow judgments of the plausibility of the model. But for early theory, when such knowledge is lacking and is precisely the goal of theory development, in tandem with empirical work, one would need a notion of plausibility that allows one to set off inquiry in a promising direction to find out about the target system. When information about the target system is not yet available, what kind of knowledge could be used to formulate plausibility judgments? Research on Bayesian models of cognition may provide useful hints.

Approaches to cognitive theorizing in the Bayesian tradition aim to characterize inductive problems, such as language learning (see Abend et al. 2017), explicitly at Marr's computational level, by focusing on the "goal" of the cognitive task and on constraints necessary to achieve that goal (Perfors et al. 2011). Studies in this tradition usually still appeal to degrees of fit of models against human data, but Bayesian methods are argued to lead to more plausible theories, because of how they push one to commit transparently to fine-grained model assumptions (e.g., properties of the hypothesis space, nature of the representations involved, how hypotheses are evaluated). The relevant notion here is *cognitive-computational plausibility*, defined in terms of explicit specifications of possible computational mechanisms (Kemp et al. 2004). Bayesian cognitive science allows considerable latitude in exploring the capabilities and limitations of models through different formal choices and model setups: degrees of plausibility can thus be introduced. For example, people have difficulty solving even basic probability problems: this is taken to suggest that Bayesian models are not cognitively plausible. However, Sanborn and Chater (2016) show that models using Bayesian sampling, instead of explicit probabilities, are cognitively plausible and empirically adequate (e.g., in explaining some limitations of human probability judgments). Another example is provided by Stenning and van Lambalgen (2010). Human reasoning is argued to be *non-monotonic*: it allows transitions between truth values or probabilities in any "direction," e.g., from true to false or vice versa, and from the prior $P(e)=0$ to the posterior $P(e)>0$. However, in Bayesian models, more evidence never makes a zero probability positive: the Martingale convergence theorems guarantee that probability distributions stabilize in the limit and require that null probabilities remain null. One could assume that probabilities are never null, only very small: but this would entail that probabilities are defined on the set of *all* propositions, not on a finite subset, which is a cognitively implausible assumption to make.

These examples show that in Bayesian theories, mathematical considerations or other *theory-internal parameters* (i.e., the way the model is set up formally), in combination with insights on what is *cognitively or computationally possible for minds and brains to achieve*, ground plausibility judgments: this is different from notions of biological or cognitive plausibility that rely entirely or largely on similarity or analogy with (unknown) aspects of the target system. Plausibility is no longer only a *relational* notion but takes on a *formal* dimension as well. This is an important shift that we explore more fully below.

Computational Plausibility: Tractable Cognition

A number of researchers approaching cognitive questions from a computational or a mathematical perspectives have made the case for an explicit concern about particular *formal properties* of theories (Barton et al. 1987; Ristad 1993). In this literature, plausibility is evaluated with reference to *computational tractability*, *descriptive complexity*, and *correlation* with human behavior. As already hinted at in the Bayesian literature, if less explicitly, attention to formalizing a cognitive problem from a computational perspective forces modelers into abstractions or idealizations (e.g., unbounded inputs or memory resources). Abstracting away from physical constraints can be considered a shortcoming of these approaches, but several authors in this framework consider instead that *formally motivated idealizations* often help avoid baking *arbitrary* assumptions into a theory (e.g., arbitrary bounds on working memory, see Savitch 1993).

At the same time, the consequent simplification of the domain of inquiry is not only desirable but necessary in the early stages of studying complex systems. We understand idealization not as contrary to plausibility, but as a requirement for plausible analyses (Szymanik & Verbrugge, 2018). On this account, simplicity of descriptions can be balanced with empirical adequacy. However, considerations on the *computational tractability* of a theory play a central role, as they enable researchers to evaluate and filter possible ("plausible") theories as they are being developed before empirical comparisons with data can be carried out (van Rooij & Baggio 2021). For instance, comparing theories of the representation of visual objects, Edelman (1997) ties criteria for plausible theories to *computability* and *space-time constraints*, arguing that "formal methods" can reveal how particular theory-internal problems may turn out to be irrelevant for the development of a broader theoretical understanding of particular cognitive tasks.

Adopting a slightly different stance here, Tsotsos (1993) suggests that no type of explanation in cognitive science is unrelated to "computational hypothesis": that is, computational considerations, in some ways, constrain *all* theories. However, Tsotsos proposes that algorithmic considerations have to go hand in hand with biological plausibility: an algorithm might be "good" (in the sense of tractability) and "valid" (accounts for experimental observations) but must also be physically realizable. Similarly, Perconti (2017) regards tractability as a computational and biological concern, while adopting a strong instrumentalist view of the goodness of theory as fundamentally tied to *empirical success*. Computational tractability does not imply cognitive plausibility; however, the former is argued to constitute "a necessary commitment" for the latter: both computational tractability and "fit to the ordinary situations [that the theories] are encoding" (empirical coverage) are required for the plausibility of a (computational) cognitive theory.

For these authors, plausibility is characterized at the intersection of satisfaction of *constraints* imposed by the *computational complexity* of the problem and the *physical resources* available for its resolution. Importantly, physical realizability may not coincide with notions of biological plausibility discussed earlier, as it is rather tied to questions about *resource capacity*, and thus relates to constraints at Marr's *algorithmic level*. In this sense, theoretical realizability can be defined within the bounds of computational complexity. Therefore, it is compatible with the limited knowledge of the system typically available during early theoretical development. So, we agree with Love (2021) that biological plausibility does not directly apply to levels other than the implementational, but we also believe that computational plausibility matters for levels other than the computational. We elaborate on this point in the next section, where we begin to articulate the value of computational insights in defining classes of potentially plausible theories.

Invariance and Tractability: Plausibility and Formal Theory

As we have seen, many of the appeals to plausibility in the philosophy of science and cognitive science collapse into other goodness-of-theory criteria. Moreover, plausibility seems needed to guide the early stages of theory development, but these other criteria are not applicable to early theory. The question is then, how do we characterize a notion that is less dependent on other criteria of good theory and appeals to considerations that do not require extensive knowledge about a target system? If the purpose of cognitive science is to provide explanatory accounts of cognitive capacities, what is desirable at the early stages of theoretical development is not to try to pursue *one* good theory, but rather to focus on the formulation of *classes of hypotheses* with high prior probability, as well as to avoid hypotheses with relatively small prior probabilities (recall Meehl), given limited knowledge of a domain (van Rooij & Baggio 2020, 2021; Bird 2021). Notions of plausibility are needed to characterize desirable attributes of early explanatory hypotheses.

Simon (1990) argued that "the fundamental goal of science is to find invariants" and that in building theories of cognition, we should aim to discover "invariants in the mechanisms that allow [us] to solve problems and learn: the mechanisms of intelligence" (p. 17). He also emphasized that these invariants will be "mainly qualitative" and "appropriate to adaptive systems" and that some will be shared "with certain nonbiological systems—the computers" (pp. 2-3). A computational lens helps navigate the space of possible explanatory hypotheses, restricting it to classes of hypotheses that are (a) *minimally plausible* (no false presuppositions or low prior probability assumptions) and (b) *invariant* across a broad range of possible algorithmic or physical implementations of a cognitive theory, and thus applicable to biological and artificial computing systems alike. Take for instance models of cognition focusing on tractability. Computational complexity theory is generally concerned with how "hard" particular kinds of problems or tasks are to solve, usually by conceptualizing them as *functions* that map problem instances to answers: e.g., the problem of deciding whether a string is well-formed given a grammar, or the problem of finding optimal outputs for given inputs based on a finite set of constraints. This perspective allows us to explore assumptions about aspects of a problem that might require various restrictions to make it tractable (Wareham 1996; van Rooij 2008; van Rooij et al. 2019).

As complexity analyses characterize *types* of problems, they also distinguish *sets* of possible hypotheses for each formulation of a problem. Consider the problem of deciding whether a string s belongs to the language generated by a grammar G (i.e., deciding whether $s \in L(G)$, the set of all strings consistent with G). There are two instances of this problem. If G is not defined a priori, we are asking how hard it is to determine if an *arbitrary string* belongs to an *arbitrary grammar* (i.e., the *universal* version of the problem; Barton et al. 1987; Wareham 1996). However, we might be interested in understanding the problem's complexity for an arbitrary string and a *particular grammar* (G is already defined). Looking at the problem in this way highlights the impact of specific assumptions about G's type: if G is context-free (Chomsky 1959), the problem is solvable in polynomial time, in both its universal and non-universal instantiations; but if G is a Lambek Categorial Grammar (Lambek 1958), the universal version is NP-complete while the non-universal one is polynomial (Pentus 2006; Heinz et al. 2009). In giving us insight into the consequences of particular problem formulations, complexity-theoretic analyses delineate *classes of desirable theories*. One can then focus on characterizing theories that are consistent with respect to some highly probable computational requirement, given a particular formulation of the problem. This is useful for generating plausible candidate hypotheses at the early stages of theory development when one typically ignores aspects of a problem that allow one to decide among fine-grained assumptions of alternative individual theories. It can also suggest how to move forward in explorations of a theoretical space (e.g., by highlighting which problem aspects should be parameterized in order to achieve minimal computational plausibility, and how; Garey & Johnson 1979; Barton et al. 1987; Ristad 1993; Wareham 1996, 1999; Heinz et al. 2009). Considerations of tractability at the computational level may also speak to representational and algorithmic commitments (e.g., they help articulating requirements for suitable data structures, given a problem specification) tying back to key desiderata (e.g., biological realism) relevant beyond the computational level (van Rooji 2008).

Delineating how classes of theories can characterize computational properties of the problem space is one perk of computational approaches beyond tractability analysis. For instance, complexity characterizations outlined by formal language theoretic analyses of string patterns have allowed researchers to establish links between linguistic phenomena and the expressivity of the machinery required to evaluate them, e.g., regular *vs* context-free strings, requiring finite state *vs* push down automata, respectively (Chomsky 1957; Hopcroft et al. 2001). Specifically, formal language theory allows for *descriptive* characterizations that address the information *necessary* to characterize a pattern of a particular class. Descriptive characterizations focus on *minimal complexity requirements* (e.g., what kind of resources are necessary to distinguish sequences of segments adjacent to each other), making it possible to isolate *invariants* of the capacities under study, i.e., necessarily applicable to *any theory* attempting to account for them as well as to a wide range of implementations of the theory (Rogers & Pullum 2011; De Santo & Rawski 2022; for experimental proofs of concept, see De Santo & Drury 2019, Bremnes et al. 2022, 2023). Complexity analyses (e.g., tractability, expressivity, generative capacity) are not meant to avoid theoretical assumptions on relevant properties of the target system. Rather, they can make such assumptions explicit (e.g., limited resources, time/space constraints), and because of this, they allow evaluation of trade-offs for classes of theories making alternative commitments (e.g., expressivity trade-offs that come with assuming trees *vs* strings as the unit of representation for language; Michaelis 2001, 2004; De Santo & Rawski 2022; Graf 2022; and references therein).

This framework also allows theorists to identify sources of complexity within the problems themselves *or* within alternative formulations of theories. Complexity considerations help isolate *invariant properties across theoretical formulations* and compare and evaluate notational variants of particular theories with respect to what each of them states about the objects of theorizing (Simon 1990; Keenan & Stabler 2010; Johnson 2015; Nefdt & Baggio 2023). This notion of invariance (under alternative *theoretical* formulations) is conceptually quite different from the invariances (under different presentations of the same or similar *patterns in data*) exploited by machine learning, as well as from other notions of invariance in science (e.g., in measurement theory; Suppes 2002). Structures or properties of a target system that are invariant in this sense are more likely to be preserved in future versions of a theory: they constitute a stable core, which is less likely to change as the theory evolves, and could also steer the theory on an early path to greater verisimilitude. Computational invariants — such as minimal complexity, fundamental limits or constants à la Simon (1990), etc. — are thus an important part of what lends a theory its initial plausibility.

As initial stages of theory development necessarily build on sparse knowledge of a system, the approaches outlined above may offer the best chances of setting us en route towards sound explanatory theories, or at least away from theories that have low prior probability or are even computationally impossible. Importantly, inferences made about a system are not just restricted to empirical, scientifically obtained data: they are also affected by contingencies of a scientific community, such as its history, common sense knowledge, and widely held assumptions (van Rooji & Baggio 2020). Fruitful notions of plausibility then should also implicitly or explicitly refer to a *community* (Agassi 2014): formal concepts of plausibility must be balanced by or filtered through a *pragmatic* model that factors in initial epistemic agreement among members of a community of inquiry.

Community and Inquiry: Logic and Pragmatics of Plausibility

A theory or claim will hardly be considered plausible if the plausibility judgment comes from only one or a few individuals, let alone if that judgment clashes with what the wider community would accept. Plausibility is a *dispositional property* of theoretical constructs that they can receive or be denied approval ("plausus") by *expert members of a community* that are expected to assess them. Aspects of the relationship between plausibility and community can be tentatively explored using tools from logic and insights from the pragmatist philosophy of science.

Belief change has been investigated in Dynamic Epistemic Logic (DEL), a modal logic to study changes in epistemic states for one or more agents.[2] In DEL, belief change for multiple agents is modeled by means of *plausibility models* (Baltag & Smets 2006, 2008), a variation of Kripke models. Intuitively, a Kripke model is a structure that characterizes knowledge or beliefs for rational agents given a set of possible worlds. A statement is *known* by the agents, if it is true in all worlds they consider candidates for the *actual* state of affairs. Agents may be uncertain about the information they possess. Their behavior and reasoning are based on what they *know* as well as on what they *believe*. If an agent

[2] Belief change has been modeled in other frameworks, like AGM (after Alchourrón, Gärdenfors, and Makinson), but DEL has become popular because of its advantages over AGM. For example, it can account for higher-order beliefs and can be applied in multi-agent scenarios. However, the epistemic and dynamic operators that enrich the DEL framework with enough expressive power to model and reason about agents' knowledge, beliefs, and actions come at a computational cost (Aucher & Schwarzentruber 2013): for instance, the satisfiability problem for individual agents in Public Announcement Logic (a fragment of DEL) is NP-complete (Lutz 2006). That said, DEL has been successfully used to model a range of problems, for example the complexity of theory of mind reasoning and related issues (e.g., van de Pol et al. 2018; Szymanik & Verbrugge, 2018).

or a group of agents *knows* something, that is true in all possible worlds accessible to them, whereas agents *believe* something if they are uncertain about it: i.e., they consider it true in some possible worlds and false in others. One can then *order* possible worlds: a *plausibility order* specifies which worlds an agent considers more or less likely to be the actual world (Velázquez-Quesada 2014). Thus, in *plausibility models*, the accessibility relation, standard in modal logic, is interpreted as a distinctive relation that reflects the agent's plausibility order over possible worlds. Suppose that an agent a is entertaining two different worlds, w and v, as possible without knowing if any of them is the actual one. They may impose a relative plausibility order denoted by $\geq_a$. To say that for agent a (or group of agents $a_1, \ldots, a_2, \ldots, a_n$, abbreviated with σ) world w is at least as plausible as v, we write $w \geq_a v$,[3] which gives a basis for comparative judgments of plausibility.

Within DEL, plausibility can be updated: it is subject to *belief revision* based on relevant information updates. One obvious way in which the information shared by a community of agents may be updated is through *public announcement*. For example, after an agent truthfully states that a proposition p in world v is false, world w might become strictly more plausible not only for her but also for other agents. Also, that p is false at v becomes common knowledge in this multi-agent system: all agents in the group know it, and know that all other agents know it.

DEL offers a structural and dynamic model of how a proposition can be deemed plausible by a community at an early stage of inquiry while remaining open for revision. Within DEL, plausibility does not presuppose that the target system is known or understood, even partly, and is not reduced to other criteria that may be more appropriate to assess theories at later stages of their development.

Modeling plausibility through the lens of DEL in multi-agent systems allows us to distinguish it from phenomenological (subjective) accounts and to appreciate its dynamic role in a community of inquiry. This leads to a notion of plausibility useful to analyze the acceptance and rejection of ideas through history, as Agassi (2014) acknowledges. In DEL, a historical perspective is introduced "for free" by the dynamic nature of formalism, allowing iterated belief revision as well as transitions from ignorance to belief to knowledge. However, a richer perspective can be achieved by incorporating key insights from post-Peircean epistemology. In that framework, plausibility can be distinguished from truth, probability, and from the background of beliefs that inquiry starts from. If hypothesis H is true, we may expect that if we conducted research on H, we would find that H would encounter no recalcitrant data and arguments (Misak, 2004). This is too much to expect when H is just deemed plausible, and no evidence exists yet for or against it: we may expect that *others would choose to conduct research* about H (rather than some H′ ranked lower in the agents' plausibility order), but not necessarily that it would survive contact with data or arguments (unless of course it is true). This pragmatic view of plausibility overlaps with criteria like pursuitworthiness, although the point raised in the "Plausibility in the Philosophy of Science" section that pursuitworthiness applies to theories at any stage of development, implies that overlap is only partial.

Pragmatists have emphasized that inquiry typically starts off from a background of beliefs that are not doubted. But what is plausible is often doubted, it must be foregrounded and *made explicit*, and need not come from the background at all. Further, probabilities often change over the short run, at each new experiment, even if slightly, whereas plausibility judgments reflect a *long-term commitment* or *policy*: researchers may be unwilling to change their plausibility order unless the truth value of relevant claims becomes known or experiments move relevant probabilities by a sufficient margin that further large shifts are not expected. In a pragmatist framework too, there is room for an epistemologically autonomous idea of plausibility, different from truth, probability, and background belief.

This point may be taken a step further. In a pragmatist analysis, plausibility can be distinguished from *verisimilitude* — a move which we are not able to make in a purely logico-inferential framework, and indeed the two ideas remain close in our treatment in the "Invariance and Tractability: Plausibility and Formal Theory" section. Plausibility judgments in a given community would be based on *consensus*: agents tend to agree on ranking H higher than suitable alternatives in their plausibility orders. Instead, judgments of verisimilitude are based on *convergence* (Misak, 2004): mathematically, two measures converge if the difference between them gradually diminishes, until it becomes so small that it can be ignored. In classical accounts (Popper, 1963, 1976; Niiniluoto, 1987), the "error" involved in theories will decrease as a theory gets closer to the truth, i.e., becomes more truth-like. The idea of *consensus*, specifically as far as plausibility is concerned, does not involve the notion of approaching a limit. In fact, it often holds even *before* inquiry starts, as the initial epistemic state of the community, and does not require that measures of error or verisimilitude are either available or applicable, as is typically the case in the early stages of theory development.

[3] Conversely, to say that for a, world v is no less plausible than world w, we write $w \leq_a v$. If w is strictly more plausible than v for a, we write $w >_a v$; if v is strictly more plausible than w, we write $w <_a v$. If w and v are equally plausible for a, we write $w \simeq_a v$ ($w \geq_a v$ and $w \leq_a v$ hold).

In DEL, one can model the distinction between initial *consensus* on plausibility judgments *vs* expected *convergence* on the truth value of hypotheses. Plausible hypotheses are the *first-order beliefs initially ranked higher* in their plausibility ordering by members of a community: e.g., that connectionist nets are plausible functional models of the brain, that Bayesian inference offers a plausible model of human inductive learning, and that initially plausible candidate functions for cognitive implementation are those with lower computational complexity. This is *consensus*: it need not be universal, and it need not remain shared as research proceeds. Agents may also have second-order beliefs about *what the community will know* as the inquiry progresses long enough. Beliefs higher in this second-order ranking capture what appears likely to obtain eventually: that is *convergence* on the truth or the probability of claims, from the agents' epistemic vantage point.[4] Importantly, only initially shared high-ranked first-order beliefs (consensus) are characteristic of plausibility judgments: agents could deem H plausible without necessarily agreeing that it will eventually be accepted as true; agents might also form first-order beliefs or commitments, and know there is agreement on them, without entertaining second-order beliefs about future shared epistemic states.

Conclusions: Plausibility and Theoretical Reform

Improving the quality of theories in any field of research requires solving a host of problems, from identifying phenomena worth explaining, to developing a set of tools for theory building and development, as well as criteria for assessing the quality of theoretical proposals. For *early* theories (i.e., theories that are not yet fully formalized, do not yet make precise qualitative or quantitative predictions, and do not encompass known facts from adjacent fields), these problems arise in particularly acute forms. Here, we have addressed one pressing problem: that of knowing whether one's nascent theory is on the right track. We have argued that the concept of plausibility — though at times used inconsistently in the cognitive science literature — may be helpful precisely at early stages of theory formation, when familiar criteria are not yet applicable, either because empirical results are sparse or inconsistent or because too little data is available to speak directly to particular theories or to support theory choice. We claim that while plausibility may not justify novel hypotheses inferentially, it provides a useful guide to their generation (e.g., leveraging analogies and other relations with more established scientific structures) and that in spite of arguments for implausible theories or models, plausibility — *properly understood* — remains a theoretical virtue.

In our view, the proper understanding of plausibility as relevant to early theory development entails at least two theses:

(1) A *formal-computational thesis* that plausible theories are those that (a) incorporate elements also shared by other proposals (*invariants*) and (b) meet minimal computational complexity requirements;
(2) A *logico-pragmatic thesis* that plausible theories are those that rational agents in a community of inquiry collectively rank higher than competing proposals at any given moment (*consensus*), regardless of whether those theories will turn out true or whether the agents believe they will.

These notions of plausibility may not be entirely epistemically autonomous from other criteria for the goodness of theory (e.g., verisimilitude plays an important role in our first thesis). Nonetheless, they do not fully collapse into other criteria, and crucially, they are applicable when other criteria are not.

Author Contributions GB, ADS, and NAN conceived and wrote the manuscript.

Data Availability No data or materials were used in this work.

Code Availability No code was used in this work.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication The work presented in this article has not been published before; it is not under consideration elsewhere, and its submission has been approved by all co-authors.

Competing Interests The authors declare no competing interests.

[4] This picture is necessarily simplified but can be refined with tools and insights that are already available in the literature. For example, the way scientific communities assess the plausibility of early theories may depend on how the members of such communities are connected among each other. Network epistemology models have shown that well-connected groups tend to arrive at a consensus quicker, but this consensus may not be correct as members can be sharing misleading evidence that might lead the community to settle on poor theory. This phenomenon is known as the "Zollman effect." Sparsely connected networks are more likely to settle on a consensus closer to the truth (Zollman 2007, 2010, 2013). Moreover, more realistic agents and interactions would need to be posited to model situations where a community is driven (e.g., by economic or other incentives) to converge on hypotheses not compatible with other criteria for good theory.

References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition, 164*, 116–143.

Achille, A., Rovere, M., & Soatto, S. (2019). Critical learning periods in deep neural networks. International Conference on Learning Representations (ICLR)

Achinstein, P. (1964). Models, analogies, and theories. *Philosophy of Science, 31*(4), 328–350.

Agassi, J. (2014). Proof, probability or plausibility. In: Mulligan, K., Kijania-Placek, K., & Placek, T. (eds) *The History and Philosophy of Polish Logic, History of Analytic Philosophy*. London: Palgrave Macmillan, London, pp. 117–127.

Aucher, G., & Schwarzentruber, F. (2013). On the complexity of dynamic epistemic logic. In B. C. Schipper (Ed.), *Proceedings of the 14th Conference of Theoretical Aspects of Rationality and Knowledge (TARKXIV)* (pp. 19–28). Chennai, India.

Baltag, A., & Smets, S. (2006). Dynamic belief revision over multi-agent plausibility models. In *Proceedings of LOFT* (Vol. 6, pp. 11–24). University of Liverpool.

Baltag, A., & Smets, S. (2008). Probabilistic dynamic belief revision. *Synthese, 165*, 179–202.

Bartha, P. (2010). *By parallel reasoning: the construction and evaluation of analogical arguments*. Oxford University Press.

Barton, G. E., Berwick, R. C., & Ristad, E. S. (1987). *Computational complexity and natural language*. MIT press.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks, 5*(2), 157–166.

Bird, A. (2021). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science, 72*(4), 965–993.

Branco, A., Rodrigues, J., Salawa, M., Branco, R., & Saedi, C. (2020). Comparative probing of lexical semantics theories for cognitive plausibility and technological usefulness. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 4004–4019).

Bremnes, H. S., Szymanik, J., & Baggio, G. (2022). Computational complexity explains neural differences in quantifier verification. *Cognition, 223*, 105013.

Bremnes, H. S., Szymanik, J., & Baggio, G. (2023). The interplay of computational complexity and memory load during quantifier verification. *Language, Cognition and Neuroscience* on-line first.

Chomsky, N. (1957). *Syntactic structures*. Mouton & Co.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control, 2*(2), 137–167.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23*(4), 305–317.

De Santo, A., & Drury, J. E. (2019). Encoding and verification effects of generalized quantifiers on working memory. *Proceedings from the Annual Meeting of the Chicago Linguistic Society, 55*(1), 103–114.

De Santo, A., & Rawski, J. (2022). Mathematical linguistics and cognitive complexity. In E. Danesi (Ed.), *Handbook of Cognitive Mathematics* (pp. 1–38). Springer.

Dror, I. E., & Gallogly, D. P. (1999). Computational analyses in cognitive neuroscience: In defense of biological implausibility. *Psychonomic Bulletin & Review, 6*(2), 173–182.

Edelman, S. (1997). Computational theories of object recognition. *Trends in cognitive sciences, 1*(8), 296–304.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. W.H. Freeman & Co.

Goudge, T. A. (1966). Plausibility of new hypotheses. *The Journal of Philosophy, 63*(20), 621–624.

Graf, T. (2022). Subregular linguistics: Bridging theoretical linguistics and formal grammar. *Theoretical Linguistics, 48*(3-4), 145–184.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11*(1), 23–63.

Heinz, J., Kobele, G. M., & Riggle, J. (2009). Evaluating the complexity of optimality theory. *Linguistic Inquiry, 40*(2), 277–288.

Hooker, C. A. (1996). The scientific realism of Rom Harré. *British Journal for the Philosophy of Science, 47*(4).

Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2001). *Introduction to automata theory, languages, and computation* (3rd ed.). Prentice-Hall.

Johnson, K. (2015). Notational variants and invariance in linguistics. *Mind & Language, 30*(2), 162–186.

Keenan, E. L., & Stabler, E. P. (2010). Language variation and linguistic invariants. *Lingua, 120*(12), 2680–2685.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. *Proceedings of the Annual Meeting of the Cognitive Science Society, 26*, 672–677.

Kennedy, W. G. (2009). Cognitive plausibility in cognitive modeling, artificial intelligence, and social simulation. In *Proceedings of the International Conference on Cognitive Modeling (ICCM)* (pp. 24–26).

Lambek, J. (1958). The mathematics of sentence structure. *The American Mathematical Monthly, 65*(3), 154–170.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Love, B. C. (2021). Levels of biological plausibility. *Philosophical Transactions of the Royal Society B, 376*(1815), 20190632.

Lutz, C. (2006). Complexity and succinctness of public announcement logic. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 137–143).

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman & Co.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*, 115–133.

Meehl, P. E. (1992a). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. In R. B. Miller (Ed.), *The Restoration of Dialogue: Readings in the Philosophy of Clinical Psychology* (pp. 523–555). American Psychological Association.

Meehl, P. E. (1992b). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological Reports, 71*, 339–339.

Meehl, P. E. (2002). Cliometric metatheory: II. Criteria scientists use in theory appraisal and why it is rational to do so. *Psychological Reports, 91*(2), 339–404.

Meehl, P. E. (2004). Cliometric metatheory III: Peircean consensus, verisimilitude and asymptotic method. *British Journal for the Philosophy of Science, 55*(4).

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 12–22). Association for Computational Linguistics.

Mewhort, D. J. (1990). Alice in wonderland, or psychology among the information sciences. *Psychological Research, 52*(2), 158–162.

Michaelis, J. (2001). Transforming linear context-free rewriting systems into minimalist grammars. In *In Proceedings of the 4th International Conference on Logical Aspects of Computational Linguistics* (pp. 228–244).

Michaelis, J. (2004). Observations on strict derivational minimalism. *Electronic Notes in Theoretical Computer Science, 53*, 192–209.

Michaelov, J. A., Bardolph, M. D., Coulson, S., & Bergen, B. (2021). Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude? *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*, 300–306.

Misak, C. J. (2004). *Truth and the end of inquiry: A Peircean account of truth*. Oxford University Press.

Nefdt, R. M., & Baggio, G. (2023). Notational variants and cognition: The case of dependency grammar. *Erkenntnis*, 1–31.

Niiniluoto, I. (1987). *Truthlikeness*. Springer.

Nyrup, R. (2020). Of water drops and atomic nuclei: Analogies and pursuit worthiness in science. *The British Journal for the Philosophy of Science, 71*(3), 881–903.

Oota, S. R., Alexandre, F., & Hinaut, X. (2022). Long-term plausibility of language models and neural dynamics during narrative listening. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*, 2462–2469.

Pentus, M. (2006). Lambek calculus is NP-complete. *Theoretical Computer Science, 357*(1-3), 186–201.

Perconti, P. (2017). The case for cognitive plausibility. In: La Mantia, F., Licata, I., & Perconti, P. (eds) *Language in Complexity*. Lecture Notes in Morphogenesis. Springer, Cham, pp. 73–79.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition, 120*(3), 302–321.

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science, 39*(8), 1824–1854.

Popper, K. R. (1963). *Conjectures and refutations*. Routledge.

Popper, K. R. (1976). A note on verisimilitude. *The British Journal for the Philosophy of Science, 27*(2), 147–159.

Psillos, S. (1999). *Scientific realism: How science tracks truth*. Routledge.

Ramakrishnan, K., Scholte, S., Lamme, V., Smeulders, A., & Ghebreab, S. (2015). Convolutional neural networks in the brain: An fMRI study. *Journal of Vision, 15*(12), 371–371.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience, 22*(11), 1761–1770.

Ristad, E. S. (1993). *The language complexity game*. MIT Press.

Rogers, J., & Pullum, G. K. (2011). Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information, 20*(3), 329–342.

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. In M. I. Posner (Ed.), *Foundations of Cognitive Science* (pp. 133–159). MIT Press.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences, 20*(12), 883–893.

Savitch, W. J. (1993). Why it might pay to assume that languages are infinite. *Annals of Mathematics and Artificial Intelligence, 8*(1-2), 17–25.

Šešelja, D., & Straßer, C. (2013). Kuhn and the question of pursuit worthiness. *Topoi, 32*, 9–19.

Shapere, D. (1966). Plausibility and justification in the development of science. *The Journal of Philosophy, 63*(20), 611–621.

Shaw, J. (2022). On the very idea of pursuitworthiness. *Studies in History and Philosophy of Science, 91*, 103–112.

Simon, H. A. (1968). On judging the plausibility of theories. *Studies in Logic and the Foundations of Mathematics, 52*, 439–459.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*(1), 1–20.

Stenning, K., & van Lambalgen, M. (2010). The logical response to a noisy world. In M. Oaksford & N. Chater (Eds.), *Cognition and Conditionals: Probability and Logic in Human Thinking* (pp. 85–102). Oxford University Press.

Stinson, C. (2020). From implausible artificial neurons to idealized cognitive models: Rebooting philosophy of artificial intelligence. *Philosophy of Science, 87*(4), 590–611.

Suppes, P. (2002). *Representation and invariance of scientific structures*. CSLI Publications.

Szymanik, J., & Verbrugge, R. (2018). Tractability and the computational mind. In M. Sprevak & M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind*. Routledge.

Toulmin, S. (1966). The plausibility of theories. *The Journal of Philosophy, 63*(20), 624–627.

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science, 69*(2), 212–233.

Tsotsos, J. K. (1993). The role of computational complexity in perceptual theory. *Advances in psychology, 99*, 261–296.

van De Pol, I., Van Rooij, I., & Szymanik, J. (2018). Parameterized complexity of theory of mind reasoning in dynamic epistemic logic. *Journal of Logic, Language and Information, 27*, 255–294.

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science, 32*(6), 939–984.

van Rooij, I., & Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry, 31*(4), 321–325.

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science, 16*(4), 682–697.

van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis*. Cambridge University Press.

Velázquez-Quesada, F. R. (2014). Dynamic epistemic logic for implicit and explicit beliefs. *Journal of Logic, Language and Information, 23*, 107–140.

Wareham, H. T. (1996). The role of parameterized computational complexity theory in cognitive modeling. In *AAAI-96 Workshop Working Notes: Computational Cognitive Modeling: Source of the Power*.

Wareham, T. (1999). *Systematic parameterized complexity analysis in computational phonology*. Ph.D. thesis, Department of Computer Science, University of Victoria, April 1999. Technical Report ROA-318-0599, Rutgers Optimality Archive.

Yang, G. R., & Wang, X. J. (2020). Artificial neural networks for neuroscientists: A primer. *Neuron, 107*(6), 1048–1070.

Zollman, K. J. S. (2007). The communication structure of epistemic communities. *Philosophy of Science, 74*(5), 574–587.

Zollman, K. J. S. (2010). The epistemic benefit of transient diversity. *Erkenntnis, 72*(1), 17–35.

Zollman, K. J. S. (2013). Network epistemology: Communication in epistemic communities. *Philosophy Compass, 8*(1), 15–27.