Computational Brain & Behavior (2023) 6:213–227
https://doi.org/10.1007/s42113-022-00166-x

ORIGINAL PAPER

On Logical Inference over Brains, Behaviour, and Artificial Neural Networks

Olivia Guest[1,2] · **Andrea E. Martin**[2,3]

Accepted: 23 December 2022 / Published online: 13 February 2023

Abstract

In the cognitive, computational, and neuro-sciences, practitioners often reason about what computational models represent or learn, as well as what algorithm is instantiated. The putative goal of such reasoning is to generalize claims about the model in question, to claims about the mind and brain, and the neurocognitive capacities of those systems. Such inference is often based on a model's performance on a task, and whether that performance approximates human behavior or brain activity. Here we demonstrate how such argumentation problematizes the relationship between models and their targets; we place emphasis on artificial neural networks (ANNs), though any theory-brain relationship that falls into the same schema of reasoning is at risk. In this paper, we model inferences from ANNs to brains and back within a formal framework — metatheoretical calculus — in order to initiate a dialogue on both how models are broadly understood and used, and on how to best formally characterize them and their functions. To these ends, we express claims from the published record about models' successes and failures in first-order logic. Our proposed formalization describes the decision-making processes enacted by scientists to adjudicate over theories. We demonstrate that formalizing the argumentation in the literature can uncover potential deep issues about how theory is related to phenomena. We discuss what this means broadly for research in cognitive science, neuroscience, and psychology; what it means for models when they lose the ability to mediate between theory and data in a meaningful way; and what this means for the metatheoretical calculus our fields deploy when performing high-level scientific inference.

Keywords Cognitive computational neuroscience · Cognitive science · Logic · Philosophy of science · Scientific inference · Metascience · Metatheory

> [A]ll science would be superfluous if the outward appearance and the essence of things directly coincided. (Marx, 1894, p. 592)

Reasoning about what our models contribute to our research is core to the computational neuro- and cognitive sciences. How do we relate the behavior of our models with psychological and neural data? In this paper, we address how common metascientific syllogisms — specifically ones that seem to be imported from the neighboring fields of artificial intelligence and machine learning — can be viewed from a formal lens. Herein, we specify and characterize reasoning in the field of cognitive computational neuroscience using formal logic in order to dissect the implications both of the reasoning itself and of what such a formal treatment grants us metascientifically.

The field of cognitive computational neuroscience, as well as its surrounding academic environs, has no doubt been radically changed by the onslaught of powerful computation, combined with the ease with which models can be constructed and applied to data, e.g., services and tools such as Keras, which provides an accessible deep learning Python library that takes advantage of graphical processing units and high performance computing services (Chollet et al., 2015). A deep artificial neural network (ANN) model — a model composed of more than two

✉ Olivia Guest
olivia.guest@donders.ru.nl

1 Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands

2 Language and Computation in Neural Systems Group, Donders Centre for Cognitive Neuroimaging, Radboud University, Nijmegen, The Netherlands

3 Language and Computation in Neural Systems Group, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

layers of individual units, which compute a summation and typically nonlinear transformation of the output of upstream units — can be created and easily trained using, e.g., back-propagation. Such an ANN can be then used as a model for brain and behavior. However, this progress in accessibility of computational tools and resources also has had ramifications for how we construct logical arguments and conceive of inference within science. This is especially true for the logical inference rules we apply in the metatheoretical decision-making processes within the cognitive, computational, and neuro-sciences, which includes determining which theories and models are useful and which are less so (cf. Rich et al., 2021). Importantly, "[ANNs are] intended to duplicate from the neural system [the] abstract computational or information processing capacity." (Chirimuuta, 2021, p. 772). To wit, given that we as a field "are [often] relying on these models as proxies for theories" (Leeds et al., 2013, p. 3), they deserve careful theoretical scrutiny. To make clear the difference between our formalization (i.e., our model) and the literature (i.e., the phenomenon), we dub our formalization of the collection of both currently formal and informal inferences rules over theories, our "metatheoretical calculus." In other words, metatheoretical calculi are proposed formalizations, i.e., models, of the way we think, created explicitly to help us reason about how we think, to facilitate communication on how we evaluate our thoughts, and to allow for improvements to both.

As we shall discuss herein, many of the same metatheoretical reasoning problems faced by the original conception of connectionism persist in cognitive computational neuroscience. As such, even though nobody disputes that connectionism has "undoubted merits" (Broadbent, 1985), the way we reason about such models will likely benefit from a comprehensive formal analysis, i.e., our metatheoretical calculus. Thus, allowing us to problematize our framings of our modelling endeavors, e.g., question what mechanistic understanding ANNs can provide. Notwithstanding, it is clear that ANNs as a modelling "framework can pave the way to new categories of scientific questions" (Barak, 2017, p. 4), provided we bear in mind that "it is not enough to know how similar a given model is to the brain: we also need to know why." (Truzzi & Cusack, 2020, p. 1).

This article joins the chorus of many other calls for better theory and metatheory (e.g., Bowers et al., 2022; Firestone, 2020; Funke et al., 2020; Jonas & Kording, 2017; Geirhos et al., 2020; Ma & Peters, 2020) — but we clarify, extend, and substantiate the argument (a) by describing, and formalizing, the discursive pattern of inferences found in the cognitive computational (neuro)sciences, by using a formal logical framework we dub a metatheoretical calculus, (b) by demonstrating how behavior, as evidenced in the literature in the form of natural language statements, when formalized can comprise a common logical fallacy, and (c) by analyzing the consequences of our metatheoretical calculus on how scientists working in the cognitive (neuro)sciences discuss and frame inferences in experimental and theoretical settings. We conclude by offering a synthesis on scientific reasoning and the desiderata for improving our inferential practice.

The Current State of Cognitive Computational Neuroscience

Cognitive computational neuroscience (CCN) can be conceptualized as the field of scientific inquiry that aims to provide "mechanistic explanations for how the nervous system processes information to support [cognition and behaviour]" (Kietzmann et al., 2019, p. 2). A mechanistic explanation involves describing how holistic properties of a complex system emerge from the causal interactions of its constituent parts (Falkenburg & Schiemann, 2019; Kaplan, 2011). To reach this goal of producing useful mechanistic explanatory theories for brain, cognition, and behavior, CCN uses various types of formal(izable) and computational techniques — both as cognitive models and as statistical tools to uncover signal within brain activity (cf., Cichy & Kaiser, 2019; Kay, 2018). Herein, we evaluate the cases in which ANNs instantiate, or stand in for, theories that furnish us with mechanistic understanding or explanation at the level of the nervous system.

A series of mainstream methodological techniques used in CCN that were developed originally by another subfield of cognitive science, specifically mathematical psychology (see Navarro, 2021; Shepard & Chipman, 1970), have shown that computing correlations over correlations can provide useful insights in terms of the structure and relationships between and within stimulus representations and between and within different organisms and models (cf., Dujmović et al., 2022). For example, we "correlate a brain region's RDM [representational dissimilarity matrix; a second-order isomorphism of internal representations] with an RDM based on one or multiple stimulus parameters (or with an RDM predicted by a computational model), [to obtain] the correlation between the two RDMs." (Kriegeskorte et al., 2008, p. 367).

Based on the discovery of such correlations over correlations, CCN proposes a theoretical position about the contents of brain states, e.g., "our IT-geometry-supervised deep representation *fully explains* our IT data" (emphasis added; Khaligh-Razavi & Kriegeskorte, 2014, p. 24), or that using ANNs "*explains* brain activity deeper in the brain [and such models] provide a suitable *computational basis* for visual processing in the brain, allowing to decode feed-forward representations in the visual brain." (emphasis

added; Ramakrishnan et al., 2015, p. 371). Furthermore, some propose that "[t]o the computational neuroscientist, ANNs are *theoretical vehicles* that aid in the *understanding* of neural information processing" (emphasis added; van Gerven & Bohte, 2017, p. 1). This betrays very strong assumptions in CCN about the explanatory virtue of correlational results and how metatheoretical inferences are drawn. This specific belief system involving correlation could be seen as the result of importing the Turing test (Turing, 1950) from computer science and philosophy of mind to CCN, without bearing in mind that the Turing test is not per se useful for furthering a mechanistic understanding, but rather for elucidating functional roles. The Turing test, in its most abstract form, evaluates if an engineered system, like a chatbot, can converse in such a way as to pass as human, i.e., can an algorithm convince a human judge that it is indistinguishable from a human? If yes, then the machine is said to have passed the Turing test and on that — functional, correlational, but not mechanistic — basis be human-like. The insights from the engineering-oriented Turing test, can lead CCN astray if we do not methodically take into account the principle of multiple realizability (Fodor & Pylyshyn, 1988; Putnam, 1967; Quine, 1951): dramatically different substrates, implementations, mechanisms, can nonetheless perform the same input-output mappings (i.e., can correlate with each other without being otherwise "the same").

Given the above discourse, what do scientists need to bear in mind when in the driving seat of these "theoretical vehicles"? How, and what do these ANNs "explain" — how, and why do they aid in "understanding"? What sort of "new framework" (Kriegeskorte, 2015) are ANNs providing us with in CCN? Calculating the correlations between our models and our empirical observations is necessary for evaluating and refining our theoretical accounts — but it is not sufficient (Roberts & Pashler, 2000) without awareness and caution when we theorize about the scientific repercussions of such modelling work. This is especially so when the focus of CCN is to hone our mechanistic understanding by "explain[ing] rich measurements of neuronal activity and behavior in animals and humans by means of biologically plausible computational models that perform real-world cognitive tasks." (Kriegeskorte & Douglas, 2018, fig. 2). We will return to this point in the "Discussion" section.

Lest We Be Hoisted by Our Own Petard

The inference rules that we deploy in the computational, neuro-, and cognitive sciences, as well as CCN specifically, to decide which theories are plausible, are supported by data, or deserve attention and consideration are often left implicit. Furthermore, such inferences can even be drawn automatically and unconsciously (Reverberi et al., 2012). By formalizing these rules into a metatheoretical calculus we can characterize transparently the mechanism by which we frame and propel our research "forwards," and by which we come to agreement about what we know in CCN. An ongoing trend in CCN, as already touched on, is using ANN models to make strong claims about what the brain, and by extension the whole human organism, is and does. From the classical connectionist approach: "[the] match between model and [human] performance [...] suggests that the representations and processes in the model *may* provide a good analog to those in the human semantic system" (emphasis added; Rogers et al., 2004, p. 229) and the model can learn the task "suggest[ing] that infant sequential statistical learning is underpinned by the *same* domain-general learning mechanism that operates in auditory statistical learning and, potentially, also in adult artificial grammar learning." (emphasis added; Mareschal & French, 2017, p. 8)

However, it must be noted that the points herein are not dependent on the waxing and waning status of ANNs within the computational sciences generally. On the contrary, these issues apply to all types of metatheorizing over formal and computational modelling (cf. Guest & Martin, 2021) that might be deploying malformed, or at least formally unexamined, inference rules. In the next few sections, we will explicate specific cases of how our reasoning within CCN is sub-par and we will typify how we (mis)use models to mediate between theory and data.

Inference Rules in (Mis)use

> [T]here are no logical forms or scientific truths in nature. Knowledge is [humanity's] construct. (Wald, 1975, p. 1)

In this section, we present a metatheoretical calculus to capture sentences as found in the wild, in the CCN literature. As mentioned, a metatheoretical calculus is a proposed formalization of the current discursive trends seen in a field of study, serving as a model of the way we think about our theories, the relevant observations, and the computational models that mediate between the two. We grant our metatheoretical calculus the right to be a formal model worthy of capturing some aspects of how we think about CCN, but, to presage what is to come, we will show how and why it reaches a paradox. Formalization this way, we propose, is a useful exercise because inter alia "sentence meanings are poised to be automatically inferentially promiscuous" (Quilty-Dunn, 2020, p. 171).

A widely deployed inference rule[1] to motivate and rationalize the use of ANNs within CCN can be readily observed in the literature as what *appears to be* — i.e., can be formally captured in a metatheoretical calculus as — modus ponens (MP). MP has the form:

$$P \rightarrow Q, P \vdash Q,$$

which can be read as: if P then Q; P is true; therefore Q is true. It is commonly deployed thus (using phraseology from Ramakrishnan et al., 2015):

> If ANNs are correlated with fMRI data, then ANNs are "a suitable computational basis" for the brain. $(P \rightarrow Q)$
> ANNs are correlated with fMRI data. (P)
> Therefore, ANNs are "a suitable computational basis" for the brain. $(\vdash Q)$

This is also the case where animals are used as models (using phraseology from Kriegeskorte et al., 2008):

> If monkey IT is correlated with human IT, then the same "behaviourally important categorical distinctions" exist in both species. $(P \rightarrow Q)$
> Monkey IT is correlated with human IT. (P)
> Therefore, the same "behaviourally important categorical distinctions" exist in both species. $(\vdash Q)$

Generalizing the syllogism, we get the following conditional to which we, as a field, apply MP:

> If the model correlates with human behavioural and/or neuroimaging data, then the model does what humans do. $(P \rightarrow Q)$
> The model correlates with human behavioural and/or neuroimaging data. (P)
> Therefore, the model does what humans do. $(\vdash Q)$

In other words, the field (by virtue of using MP) is asserting (based on what is presented as empirical evidence) that P is true. We, as a field, are also asserting that the conditional $(P \rightarrow Q)$ is a useful and/or verisimilar formulation of what we believe about the brain and behavior (see Fig. 1a). To presage the next section, this is in fact a type of "pizza problem" (Guest & Martin, 2021) — while superficially formal(ized) and seemingly sensible, it is in fact unfounded and goes against our expressed empirical and theoretical understanding of cognition.

Additionally, by the same token we have granted ourselves the ability to express ourselves and derive truths using MP, we also (by definition) could deploy modus tollens (MT); see Fig. 1. MT has the form:

$$P \rightarrow Q, \neg Q \vdash \neg P$$

Thus, in this case, MT (see Fig. 1b) would take the form:

> If the model correlates with human behavioural and/or neuroimaging data, then the model does what humans do. $(P \rightarrow Q)$
> The model does not do what humans do. $(\neg Q)$
> Therefore, the model does not correlate with human behavioural and/or neuroimaging data. $(\vdash \neg P)$

For example, we could deploy MT in the case of ANNs' visual object recognition diverging greatly from people's, e.g., when ANNs encounter (i.e., when we engineer) adversarial images as input (Szegedy et al., 2014). Adversarial images are collections of pixels that do not look at all to a human observer like the class label(s) returned by the ANN. They typically look (to a human) like totally unrelated images of scenes or objects. Something encoded in the pixels, but imperceptible to humans, has been perturbed and so, e.g., a photo of what is "obviously" a panda is classified as a gibbon by the ANN; or, e.g., an image of something abstract-looking (like stripes or some other texture or repeating pattern) is classified as a specific object or scene (for more examples, see Dujmović et al., 2020).

When the models, in such adversarial cases, fail to classify images like a human, we do not conclude that this makes ANNs by definition unhuman-like. We do not construct this MT-based syllogism, even though nothing explicitly stops us since we happily deploy MP above (see Fig. 1):

> If the model correlates with human classification on photorealistic stimuli, then the model is impervious to adversarial images. $(P \rightarrow Q)$
> The model is not impervious to adversarial images $(\neg Q)$
> Therefore, the model does not correlate with human classification on photorealistic stimuli. $(\vdash \neg P)$

Instead, we tend to conclude that either the way the ANN has been trained, or otherwise designed, is dramatically different to humans (for example, Dujmović et al., 2020; Geirhos et al., 2020: Linzen & Leonard, 2018) or alternatively that indeed there is potentially something human-like about the (mis)classification of adversarial images (for example, Elsayed et al., 2018: Zhou & Firestone, 2019). In other words, instead of MT, scientists claim that ANNs that diverge from human performance need only to be modified somehow. They need to be further "aligned" with brain and behavior data (Peterson et al., 2016) and/or they need to further "incorporate" cognitive mechanisms (Luo et al., 2021). Thus, it is widely

[1] We embed the arguments within first-order logic as we interpret them from the literature; we do not advocate applying deductive inference rules to problems of induction.

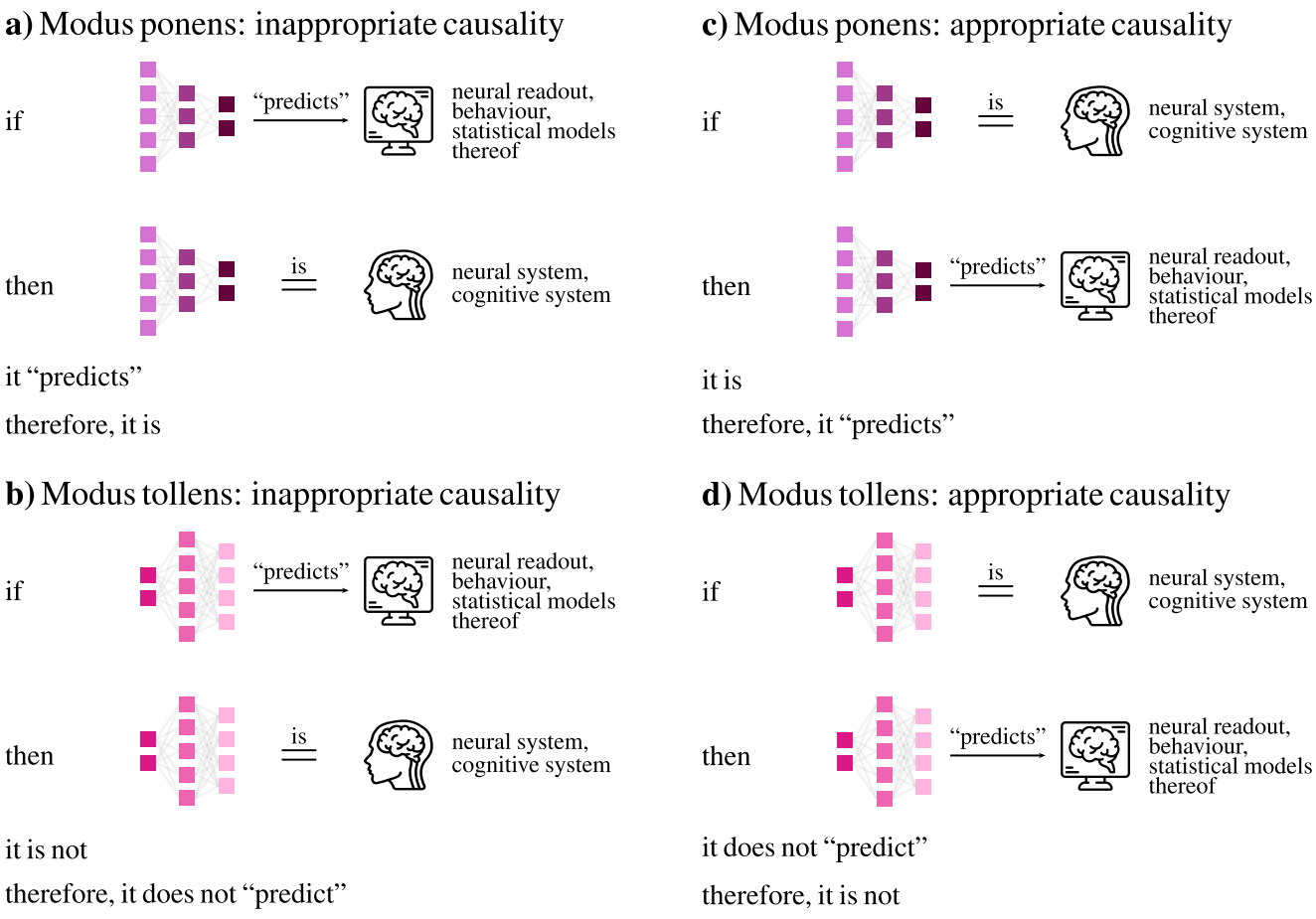

Fig. 1 Visual depictions of modus ponens and modus tollens applied over both inappropriate and appropriate causal relationships between models (represented by boxes and lines to denote artificial neural networks) the cognitive and neural systems the models try to capture (represented by a human head and brain) and the data collected from such systems (represented by a brain on a computer monitor). On the left, the degenerate syllogism found in CCN which superficially resembles modus ponens (panel a, top left) and of the mirror-image but largely unused variant, which would resemble modus tollens (panel b, bottom left). On the right, the causal relationship as it actually stands with modus ponens (panel c, top right) and modus tollens (panel d, bottom right) applied. (Credit: Icons designed by Smashicons from Flaticon)

accepted that "to achieve human level performance, [such models] will need to [incorporate] characteristics of natural intelligence" (McClelland & Botvinick, 2020, p. 25). Once "updated" in these ways, ANNs will, for example, "not be subject to adversarial [images] that seem so bizarre to humans, and will show the same set of strengths and weakness[es] (visual illusions) that characterize human vision." (Dujmović et al., 2020, p. 13).

We often entertain models that can do things that humans cannot. For example, we employ models with superhuman memory or that can learn statistical dependencies that are outside the scope of human perception (viz., all ANN, but especially deep, recurrent, and convolutional architectures can learn beyond human capacity in certain circumstances and tasks, ergo the logic and benefit of applying such systems in machine learning). But we do not take this capacity as evidence that the model is failing to approximate human behavior. Similarly, any model that can reproduce a pattern of neural activity is likely to be able to produce a pattern of activation that the human brain does not or cannot produce. Yet, this inconsistency is not an impediment to our field's logical inference practice. Even though Q can, and often does, fail to be true, we, as a field, do not formulate its relationship to P in terms of MT. This is prima facie untenable — a heightening of contradictions within CCN's metatheoretical calculus — given the rules of formal logic. If MT is dis-preferred, predominantly avoided, treated similarly to how negative evidence is treated in scientific inference, why is MP accepted?

Affirming the Consequent

Herein, we have presented a metatheoretical conditional statement that we in CCN subscribe to:

> If the model correlates with human behavioural and/or neuroimaging data, then the model does what humans do. ($P \rightarrow Q$)

The structure of this argument is inappropriate in two important, related ways. First, it is inappropriate because we propose that nobody explicitly believes this about complex systems like the brain. So even though the CCN literature often deploys MP based on this, the conditional is *not* how we, as a field, evaluate models more broadly. Correlation of our models to data, i.e., a good fit, is necessary but not sufficient (Roberts & Pashler, 2000). CCN is an outlier in terms of the centrality of seemingly malformed premises and ill-posed arguments within the rhetoric provided to support why ANNs are useful models. In other words, models — in cognitive science generally — are evaluated with more metatheoretical awareness than merely checking if they correlate with data, or have the highest r-squared. Theoretical contributions definitionally must be evaluated not primarily on their predictive power (which makes sense for statistical data models), but on their explanatory virtue. Computational and/or formal models in cognitive science are indeed often juxtaposed to a swathe of empirical evidence to show they can recapitulate behavioral or neuroimaging data (Guest & Martin, 2021). But imagine if we only evaluated cognitive models based on the amount of variance explained (viz., r-squared or AIC/BIC), or if correlation was the only criterion for identity? How many things would we confuse with the brain or as the arbiter of behavior (cf. Meijer, 2021, for an example with cryptocurrency and rodents)? Thus, inferences to "models doing what humans do" (i.e., our Q) based on such correlations are not permitted due to lacking theoretical and empirical support. In other words, as we shall explain, $P \rightarrow Q$ is a problematic construction if it is not explicitly tethered to or embedded in the context where the inference is taking place.

The nature of our inferences can be improved if we take a few steps back and consider our theorizing before asserting correlation is a stand-in for causation or explanation. For example, consider what "explain" means in these extracts: "computational models from computer vision and neuroscience can *explain* the [inferior temporal cortex] representational geometry in human and monkey" (emphasis added; Khaligh-Razavi & Kriegeskorte, 2014, p. 23) or "intermediate model layers best *explain* primary auditory cortical responses, while deeper layers best explain voxels in non-primary areas." (emphasis added; Kell et al. 2018, p. 631). Correlation is typically highly conceptually distinct from explanation, but here it is identically used (cf., Cummins, 2000).

This brings us to the second reason that the conditional presented is malformed: it is not an appropriate description, model, of the empirical evidence and theories we have at our disposal. Thus, it neither describes the status of (meta)theoretical claims we, as a field, make with respect to models (i.e., the high-level calculus we use to evaluate theory), at least outside CCN, *nor* is it backed-up by any evidence. Importantly, when we speak about our scientific findings we have to do so in ways that are consistent with our field's beliefs and assumptions. Alternatively, if we disagree with the beliefs and assumptions of our field, we must do so explicitly and in a clear and transparent way.

Based on the above, we in CCN are implicitly affirming the consequent within the metatheoretical calculus we have provided. The proper relationship between P and Q is the converted (i.e., the order is swapped) conditional to that which is described (see Fig. 1c and d):

> If the model does what people do, then the model correlates with human behavioural and/or neuroimaging data. ($Q \rightarrow P$)

What we previously called MP is not MP — it is affirming the consequent:

> The model correlates with human behavioural and/or neuroimaging data. (P)
> Therefore, the model does what humans do. ($\vdash Q$)

If we want to computationally model in CCN we could explicitly propose: if our model is capturing something mechanistic (Craver & Kaplan, 2020; Kaplan & Craver, 2011), as well as its functional role, about the brain and behavior (Q), then we collect evidence (i.e., correlate the model with empirical observations) to test, support, and improve our model (P). Converting, flipping the order of, this manifestation of the conditional demonstrates not only that $P \rightarrow Q$ could lead to a fallacy, but also highlights that it is unlikely that we affirm the consequent so brazenly in other, broader, scientific contexts (compare the four panels in Fig. 1). In other words, $Q \rightarrow P$ highlights the metascientific relationship between P and Q. The "sense" of Q is not contained in the "sense" of P (Sundholm, 1994) — but vice versa. The potential presence of this fallacy in, or of readers automatically drawing this inference from, influential papers in the literature indicates a likely confusion between types of inference (cf. Blokpoel et al., 2018), a misunderstanding of the evidentiary role correlation plays, and a lack of formalized thought on the relationship between model and observation.

In this way, ill-posed argumentation is unwittingly permitted during (meta)scientific inference in CCN. The "state of affairs" in CCN does not "obtain," i.e., it can never be a true statement about the world (Sundholm, 1994). That is to say, the empirical universe that we collect observations from is not set up, as far as we know, to support $P \rightarrow Q$; see Table 1, for a synopsis of the authors' claims. And so

Table 1 The authors make the following core claims about modelling work in CCN

Key concept	**Function**
Obtaining, the property of a syllogism or argument, when expressed within a formal system, to be true* given the particular world the argument it is situated in, is a vital property of metatheorizing and relating models to phenomena.	Whether our metatheoretical calculus (the relationship we propose between models and observations) obtains is a function of how we relate our models to the cognitive and neural systems we wish to understand and describe.
Impediments to inference	**Proposed solutions**
The literature explicitly or implicitly applies deductive argument forms to inductive reasoning; the relationship between model, explanandum, and explanans is inconsistent or obscure.	Focus on precise formulation of the relationships between models and the phenomena they are trying to capture, i.e., focus on obtaining.
ANNs are useful models for the brain, behavior, and cognitive capacities; both as proof-of-concept models, and as mechanistic models.	In terms of mechanism, ANNs may or may not resemble the cognitive and neural systems that their behavior or performance approximates. However, models must relate to the phenomenon they purport to explain in ways that obtain. The models resemble the phenomena because they indeed somehow capture (our beliefs about) the essence of the phenomena, and not vice versa, i.e., models are not capturing the essence because they are correlated with the phenomenon.
We often deploy a metatheoretical calculus that constructs syllogisms or arguments that do not obtain in CCN. Models often relate to their phenomena in ways that do not respect the causal structure we already commit to as a field, will never obtain. When the causal arrow that mediates the relationship between model and phenomenon is the wrong way round, the empirical universe cannot be truth-making for that relationship, by definition.	We must ask ourselves: • do our beliefs about complex systems, such as the principle of multiple realizability, ever make our metascientific calculus true as described herein? • is the world populated by cognitive and neuroscientific observations set up to be truth-making for the causal relationship between a given model and the world? The way to rectify the relationships between models and phenomena, e.g., between ANNs and cognitive capacities, is to express the relationships explicitly and with directionality. Additionally, probabilistic inference is subject to the same constraints, because it has the same stipulations about how causal relationships may obtain, i.e., if the empirical universe is not set up to ever make their premises be true.

*The meaning of truth here is merely formalism within the given system, in our case formal logic. This is not to imply any notion of truth other than that fully inside a formalized system

the literature containing these "curious shadowy" (Russell, 1918) syllogisms will never obtain, i.e., will never make $P \rightarrow Q$ a verisimilar proposition, description, of the causal relationship — it will always be falsified. The only way we can envisage a state of affairs in CCN obtaining is if we explicitly commit to $Q \rightarrow P$ (recall Fig. 1). This issue is not uniquely explainable by claiming that we do not know how to use formal logic, or specifically that we affirm the consequent. A big part of this, we propose, is a loss of clarity between materially licensed (Norton, 2003) induction, using what we know about computers and cognition, and inferences, including inductive ones, which otherwise do not obtain (Sundholm, 1994).

How We Fall(acy)

> And if thou gaze long into an abyss, the abyss will also gaze into thee. (Nietzsche, 1886)

Metatheoretical Considerations

Typical (meta)theories in CCN take on forms such as "the brain does what the ANN model does because the ANN model was trained on the same type of data as the brain" or "cognition works this way because the ANN model learned to approximate task performance." Through this, we propose CCN permits, leaves the door open to, a logical fallacy — affirming the consequent — to be deployed when interpreting computational modelling successes (MP), but not failures (MT).

We, the authors, wish to warn against overextending, or indeed wrongly deploying, these types of syllogisms. To close the door on this possibility, we must scrutinize how we discuss our work. "Although these models have been developed with engineering goals rather than *neurocognitive plausibility* in mind, recent neuroimaging studies have shown a remarkable correspondence between the layers of [ANNs] and activation patterns in the visual

system." (emphasis added; Devereux et al., 2018) This is not a fallacy; it is a statement about the current state of research in CCN. However, overextending the above can result in assuming that the close match between model layers and observations from imagining the brain implies "neurocognitive plausibility" — a phrase commonly used to mean that the model mechanistically, not just functionally, matches the brain. Similar arguments, open to overextending, can be found in many sub-areas of CCN, such as that ANNs are "a novel *biologically feasible* computational framework for studying the neural basis of language." (emphasis added; Goldstein et al., 2021) The problem is that we do not yet know, or agree on, what the brain's mechanisms are — this is the stated goal of CCN — and so we cannot claim that something is more or less "plausible" without conceptually engineering (Chalmers, 2020; Love, 2021) "plausibility", thus shutting the door to it functioning as a weasel word (Jason, 1988).

Importantly, CCN does not only deploy neural networks as (neuro)cognitive models, but also uses ANNs as black box models (cf. Kietzmann et al., 2019), performing in both cases similar (meta)theorizing — for example, "[t]he fact that recognizable features of stimulus images could be reconstructed with a simple linear model [what we have generalized to statement P in this paper] *indicates* that the latent space represents properties that are also represented in brain activity [Q]." (emphasis added; Seeliger et al., 2018, p. 781). Similarly, "computer vision recognition systems *may* serve as viable proxies for theories of intermediate visual object representation." (emphasis added; Leeds et al., 2013, p. 1). Notwithstanding, there is more mindful phraseology in the broader connectionist literature, e.g., "[t]he close match between model and [observations] suggests that the representations and processes in the model *may* provide a good analog to those in the human semantic system" (emphasis added, Rogers et al., 2004, p. 229); as well as in CNN, e.g., "[t]he categorical and continuous aspects of the representation are both consistent between man and monkey, suggesting that a code common across species *may* characterize primate IT." (emphasis added, Kriegeskorte et al., 2008, p. 1138). The use of "may" makes the syllogism probabilistic modus ponens in a clear way, which might not leave the door open to accidentally affirming the consequent. Importantly, however, this is only true in a fully formal setting and natural language can still lead to affirming the consequent (Collins et al., 2020; Collins & Hahn, 2020; Quilty-Dunn, 2020; Reverberi et al., 2012). Avoiding affirming the consequent can also be subserved by the clarification that ANNs "may simply rely on brute-force memorization and interpolation to learn how to generate the appropriate linguistic outputs in light of prior contexts" (Goldstein et al., 2021) — something which does appear to be true in certain contexts (Zhang et al., 2016). Others solve this by carefully couching their findings as explicitly correlational where applicable, hand-in-hand with conceptually analyzing the capacity under study (e.g., Lindsay & Miller, 2018; Nicholson & Prinz, 2020). Either way, leaving our syllogisms ambiguous — this includes not explaining that it is $Q \rightarrow P$ and not $P \rightarrow Q$ — leads to, or at least does not protect from affirming the consequent. The consequent and antecedent in our writings must be explicitly "the right way round" exactly because confusion exists.

Relatedly, data from the brain is not inherently mechanistically informative in and of itself, keeping in mind that CCN is about uncovering mechanisms. Just like behavioral data, neuroimaging and other types of brain data (e.g., single-cell recordings) do not constitute a mechanism. However, such data, as holds for all types, can indubitably aid in adjusting our mechanistic understanding during theory building. An understanding of mechanism is gifted to us by theoretical proposals for such mechanisms, for which we then collect evidence. Recall it is "if the model captures human capacity, then the model approximates human data" ($Q \rightarrow P$; see the "Affirming the Consequent" section) and not vice versa — causation implies correlation and not vice versa. These types of data, in fact all data, support (or not) our theoretical proposals, but they do not constitute them. Data are useful for building theories, but they are not sufficient in and of themselves to form a theoretical account. Data is collected with the intention to support some theoretical position, and thus is imbued with theoretical assumptions. However, and for that reason, data are not identical to a theory. $Q \rightarrow P$ obtains, while $P \rightarrow Q$ does not, and results in a fallacy when modus ponens is applied.

Another possibility is that a common belief in the field could be that a model approximating (through multiple extensions or expressions, including model behavior, processing, or output) or quantitatively predicting (again via multiple extensions) human behavior or neural data is equivalent to "providing an explanatory model." This has in it the spirit of modelling natural phenomena in order to better understand them; however, the question then becomes, what would that mean for an ANN to be an explanatory model of brain computation? If one thinks that an ANN (and its behavior, performance, or processing patterns) constitute an explanatory model of "how the brain does x," then what does that belief entail? If we charitably assume that what our field considers to be explanatory are mechanisms that are extensions of the causal structure of a natural phenomenon (a la Kaplan & Craver, 2011; Craver & Kaplan, 2020), then, it implies that thinkers in the field believe that some essential property, or properties, of ANNs explain how the brain does x because they serve as mechanisms. What could those properties be? How are they mechanistic? To begin to answer both questions, we would encourage the field to name these properties explicitly and conceptually analyze

whether they constitute mechanisms. For example (personal communication, K. Srinivasan, R. Ajemian, R.C. Berwick. January 14, 2022): Contemporary deep learning ANN typically form tessellated linear mappings, or diffracted hyperplanes, between input and output spaces. Diffraction makes the mappings appear nonlinear, and the degree of tessellation is a function of the input space and the depth or number of layers. Functions like RELU or softmax on outer layers force discontinuity on the output space, thereby creating the non-linearity between input and output space. In this sense, the output is a nonlinear transformation of the input space. By analogy, is the claim then that how the brain does a computation, or "does x" is by performing a nonlinear transformation of an input? That must be true. But, because this generic statement (just a description of the defining properties of a deep ANN) indeed describes everything the brain does, it does not explain what the brain computes, nor how it does so. It is sadly rendered trivial. In other words, if ANNs are to ever offer a mechanistic explanation of brain computation, much work must be done to determine whether that mechanism can be anything other than logistic regression.

Functionalism Versus Mechanism-ism

What *do* "mechanism" and "functional role" mean in the context of modelling in CCN, and in neuro-, cognitive computational sciences generally? Importantly, we do not doubt that all relevant types of data do indeed give us "a key opportunity to test [and refine] several planks of the deep learning hypothesis" (Saxe et al., 2020) or indeed any relevant theories or hypotheses. However, there are other "planks" (as Saxe et al., 2020, also note). How that data is used metatheoretically matters as much as, e.g., what statistics we can compute between and within observations and ANNs. In other words, we have to form metatheoretical syllogisms that involve this data and their relationship to our models. Analyzing and understanding these syllogisms is key to doing science openly, coherently and consistently — we shut the door to formal logical fallacies. Let us, therefore examine the reasons why $P \rightarrow Q$ does not obtain.

Functional role can be seen as the high-level description in terms of how inputs are transformed into outputs, e.g., describing the capacity of children to perform addition as "children perform addition." *Mechanism* is the way in which a function is implemented in a physical substrate, e.g., humans can perform addition mentally or using a calculator. Functional role and mechanism are confusing without a given level of analysis or context. For example, "a computer programme has functional transparency if it is possible to know the algorithm that the programme instantiates" (Chirimuuta, 2021, p. 780). However, that is not how code works in practice at lower levels of analysis. Much like when understanding cognition we might not understand what individual neurons are doing, we do not (need to) know how our, e.g., R code is turned into machine code and what algorithms the hardware is using — mutatis mutandis for what algorithms the electronics below the hardware level is using, and the physics below that, etc. For example, speculative execution (Lampson, 2006) means that as a user, we do not actually know what CPU-level algorithm is being used to instantiate our code-level algorithm. Speculative execution, present in modern CPUs, predicts which instructions, e.g., post conditional branching, could be needed in the near future and executes them. This by definition means that the actual algorithm carried out by the hardware may be different to what we might think, given branches are treated in ways that might not be obvious to the programmer unaware of speculative execution at the CPU-level. Such principles from computer science and engineering can be, if done carefully, imported into how we carve neuroscientific nature at its joints. That is to say, functional transparency is a very useful concept once we delimit the context or layer of analysis of interest (cf. Kaplan, 2011; Potochnik & Sanches de Oliveira, 2019; Zipser & Andersen,1988).

A Broken Metaphor Is Right Twice a Day

> [A] single connectionist model can simulate results that imply mutually exclusive psychological processes. Thus, results consistent with a connectionist model should not be taken as evidence for the model. [ANNs] can retard the discovery of the information that a subject uses in a task. (Massaro, 1988, p. 219)

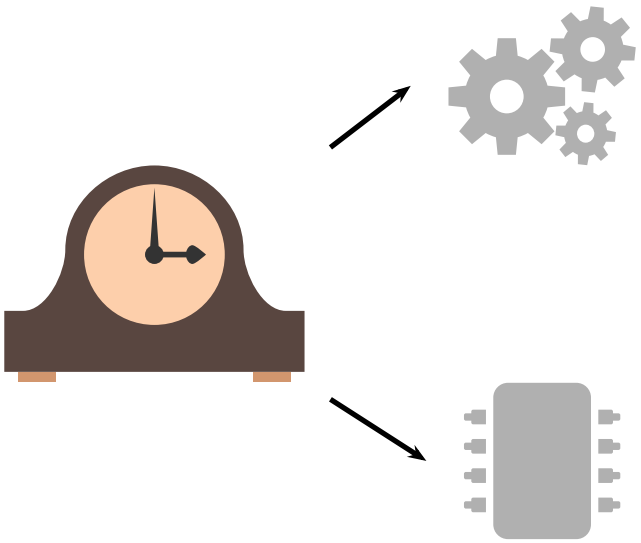

Fig. 2 A clock can be "behaviorally" (i.e., externally) identical to another clock, completely independently to each clock's implementation. This serves as a very simple example of why multiple realizability is important when understanding complex systems. (Credit: The mantelpiece clock icon is by EmojiOne (CC BY SA 4.0). The cogs icon is by Team Redux (Open Font License). The microchip icon is by Dave Gandy (CC BY 4.0))

In Fig. 2, we can see a visual depiction of a variant of the same fallacy we have described herein. Let us frame this example two different ways. First, there are two clocks with identical appearances (as shown on the left in Fig. 2) and functional roles: a mantelpiece clock that indicates the time of day. However, the two clocks are implemented using dramatically divergent mechanisms: clockwork (top right) and digitally (bottom right). This exemplifies a typical case of multiple realizability, both types of mechanism (can) give rise to identical behaviors, while having no mechanistic similarity. One can learn about the time of day by looking at either clock, but one cannot learn about the internals of the one clock by looking at the other "wrong" clock.

Ignoring the principle of multiple realizability, risks (if not ensures) confusing the explanandum (what we are attempting to explain; human cognition) and the explanans (our explanation; our model). Recall, our degenerate syllogism: if the ANN model behaves like human cognition, then the model is cognition. Importantly, it might be the case the model is exactly like human cognition, but we cannot safely conclude that from the premise given multiple realizability. The same goes for the clock in Fig. 2, we can, for example, set the explanandum to the digital clock and the explanans to the clockwork clock. If we apply the same syllogism, we get: if the clockwork clock behaves like the digital clock, then the clockwork clock is a digital clock. This is deeply problematic and readily falsifiable — mutatis mutandis for vice versa. For more directly scientific confusions between model and phenomenon, explandum and explanans, consider the motion of objects under gravity and Newtonian mechanics. $P \rightarrow Q$ in this case, takes the form: if Newtonian mechanics behaves like physical objects, then Newtonian mechanics is physical objects. None of these syllogisms obtain exactly because models can be multiply realized, and because models are different qualitatively to phenomena.

To really hammer home how multiple realizability relates to our modelling case in CCN — specifically how it demonstrates that the reasoning in use is flawed — we propose a complementary, augmented way of looking at Fig. 2. Instead of seeing two possible options for instantiating a clock, we can conceptualize one clock, e.g., the clockwork one, as the real empirical clock and the, e.g., digital clock, as the "computational" model. In such a scenario, it should be readily obvious that conclusions with respect to understanding the clockwork mechanism of the "empirical" clock, bar behavior, cannot be safely drawn by looking at the mechanisms. Importantly, unlike the idealized engineered clock example we do not know in non-engineered complex systems, like the brain, what the specification of the behavior is with full certainty. We merely can take a view supported by our theory-laden data — we cannot ever know the "ground truth" of the brain's specification in the same way we can of a timepiece. Therefore, parallels between mechanisms, also known as substrates or implementations, even when they give rise to identical behaviors, e.g., clockwork and digital clocks, cannot be safely drawn without further constraints.

(Un)licensed Analogies

All modelling, especially theoretical, cognitive, neuroscientific modelling, as found in CCN, can be seen as drawing analogies between models and empirical phenomena. However, to draw inductive inferences (like argument from analogy) from models to phenomena under study to organisms under study and back we must do so cautiously, transparently, and perhaps most importantly — since induction has no "universal schemas" (Norton, 2003) — in a way traceable to agreed upon facts. That is to say, if we know the melting point of one sample of a chemical element, we can generalize this knowledge to all instances of that element. In contrast, if we know the melting point of one piece of wax, generalizing to all wax is not licensed. We are licensed in the first case, but not the second, not from the logical form of the inference itself, but "from facts pertinent to the matter of the induction" (Norton, 2003, p. 4): a chemical element is taken to be homogenous, while wax can be composed of various substances with differing melting points.

Turning our attention back to the specific type of induction relevant here, argument from analogy: what does this mean for our conditional "if our model correlates with neurocognitive data (P), then our model can do everything else the human organism does (Q)"? Argument from analogy has the form (Salmon, 2013):

> X has properties α, β, γ, etc.
> Y has properties α, β, γ, etc., and also an additional property ω.
> Therefore, X has property ω as well.

To demonstrate our $P \rightarrow Q$ is a false analogy, recall the example in Fig. 2 with the clocks. We can reformulate "if a clockwork clock behaves like a digital clock (P), then the clockwork is the same as a digital clock (Q)" to the following false analogy — an induction that is not licensed by the facts:

> Digital clocks display the time.
> Clockwork clocks display the time, and require manual winding.
> Therefore, digital clocks require manual winding.

This is an example of an argument from analogy, which must be licensed by material facts if we want it to locally obtain (Norton, 2003). In the case of CCN:

> ANNs correlate with fMRI data.

> Brains correlate with fMRI data, and instantiate the biological mechanisms for cognition.
> Therefore, ANNs instantiate the biological mechanisms for cognition.

Both these inductions for clocks and brains are not licensed because the principle of multiple realizability makes it unsafe to argue from analogy because it is the mode through which the conditional becomes false, and the argument rendered unsound.

Importantly, recall how we correctly avoid MT in CCN, we realize the analogy between model and human somewhere has become scientifically useless. When a model fails to categorize an adversarial image, we say the similarity between it and humans has ended, and not that ANNs are holistically so different to be completely inappropriate models. We notice our $P \rightarrow Q$ is unsound (the premises false) because we rested on an inductive inference that was not licensed, i.e., a false analogy. Thus, within CCN when we make metatheoretical decisions we should be able to answer for ourselves: is this induction licensed, is this a sensible analogy, given multiple realizability, given that we do not (yet) know the mechanisms of the brain?

When I Think About You, I Adjust My Calculus

Affirming the consequent is a repercussion of entering a universe that is not truth-making for our inference $P \rightarrow Q$: what is described as a useful inference from models to brains does in fact not obtain. Nonetheless, it is likely the case that when we invite, if not commit, the fallacy, we do so because we actually believe it is MP, not because we are aware it is a fallacy. Thus, an alternative explanation for why our field has the propensity to permit minimally accidentally hinting at the fallacy is that it is forced to operate in a world where no explanation obtains. In other words, one cannot make inferences to the best explanation, or abduction, if every possible choice for explanation does not (currently) obtain. Abduction only works if we have a selection of good explanations to work from. We may simply not yet be in such a state of affairs in CCN. However, we may be able to arrive at such a state — and importantly, abduction has the same logical form as affirming the consequent (Frankfurt, 1958; Plutynski, 2011).

The authors oppose prescriptivism as a remedy to our afflictions, logical or otherwise. Notwithstanding, we propose the following steps for consistent reasoning in CCN, to adjust our syllogisms, both internally (to avoid affirming the consequent) and externally (to obtain). (a) Create a metatheoretical calculus explicitly of our projects or papers and evaluate the logical consistency of our premises and arguments, especially keeping track of causal, e.g., versus rhetorical or temporal, relationships between model and observation; and bearing in mind that theories imply data, and not vice versa (i.e., the Duhem–Quine thesis, Harding, 1975). (b) Explicitly discriminate between functional role and mechanism during theorizing, taking context and multiple realizability into account. (c) Couch our metatheoretical calculus in terms of abduction over the path function such that it obtains. The path function from (Guest & Martin, 2021) can serve as a rudimentary basis for all this if a (fully) formal approach is not (yet) beneficial, as can the questions and example answers in Table 2.

Discussion

Herein, we propose that our interdiscipline, CCN, needs to (re)evaluate the contribution of the new flavor of connectionism that purports to use deep ANNs to derive and refine our theoretical explanations and understandings of the brain (cf. Chirimuuta, 2021; Guest et al., 2020; Martin & Doumas, 2019, 2020). To this end, a series of core questions need to be asked when we carry out such work, as shown in Table 2.

Table 2 Questions and example responses we could ask and answer for ourselves when engaging in metatheorizing, thinking about what our modelling work contributes, within CCN

Potential question	Example answer
How does the model mediate from theories to data and back again?	It allows us to test if our assumptions, especially simplifying assumptions, about cognition and neural systems can lead to behavior, including internal representations, that correlate with those of target natural systems.
How do our field's models bridge levels of analysis, and are these bridges scientifically useful?	The model captures important aspects of function (based on these tasks, neural readouts, etc.) and mechanism (based on these proposed causally interacting components).
What do the models we develop in CCN offer in the context of psychology, cognitive science, and neuroscience in general?	The model allows us to explore the repercussions of our assumptions about representations and input-output mappings, especially ones that correlate with those of our observations from the brain, cognition, and behavior.

The way we have operationalized the issues found in the literature currently indicates the possible misapplication of logic resulting in affirming the consequent. Alternatively, this might be due to a deeper difference in ideological or ontological perspectives with respect to what CCN aims to uncover, how mechanism is seen by CCN scholars, or indeed how success of a model is interpreted, known as the success-to-truth inference (Wray, 2013; Vickers, 2019).

Cognitiva (Meta)scientia Redux

> [T]he extent of the match between a model representation and the neural data is appraised solely based on the correlation between the empirical dissimilarity structures constructed with neural recordings and model representations. (Tacchetti et al., 2017, p. 11)

Misapplication of such metatheoretical logic, as we have expounded on herein, contributes to overpromising and underdelivering, impeding the progress of CCN as much as it does the fields that touch on AI generally. We do not argue that ANNs do not make highly useful models within CCN and the neuro- and cognitive sciences more broadly. Although "machine learning provides us with ever-increasing levels of performance, accompanied by a parallel rise in opaqueness" (Barak, 2017, p. 5), we do not believe using machine learning this way is the only source of lack of transparency and of lack of "open theorizing" (Guest & Martin, 2021) within CCN. Interdisciplines, like cognitive science, strive to properly allow for the scientific exchange of ideas and methods within and between their constituent participating disciplines. We wish to facilitate dialogue on how to theorize usefully when importing ideas to CCN from other related fields. An example of a maladaptively imported idea is that of the Turing test (Turing, 1950), which involves understanding and contextualizing the principle of multiple realizability. The Turing test sets out to differentiate the human(-like machine) from an algorithm, essentially a chatbot. Through behavioral probing, e.g., asking questions in natural language, the person performing the test attempts to ascertain if the agent (machine or person) answering is responding meaningfully differently to a person. If the machine can "trick" us into thinking it is a human, it is said to have passed the Turing test. The Turing test is very useful if we want to engineer algorithms that can exchange details with people seamlessly. However, if we take this test and use it to infer more than perhaps Alan Turing intended, that the machine indeed is a person (our $P \rightarrow Q$), we have slipped into affirming the consequent and false analogy.

As we have shown, it is not unusual if formally treated to discern or derive fallacies in the CCN literature such as cum hoc ergo propter hoc (i.e., with the fact, therefore because of the fact, or "correlation does not imply causation"), begging the question, confirmation bias, false analogy — the root of these informal fallacies is the formal fallacy of affirming the consequent. The Turing test and similar functional role-based analogies allow us to stumble into a formal fallacy if we stray far from engineering systems and towards understanding human cognition. Ultimately, the lack of attention to the high potential for (mis)application of formal logic in CCN betrays its current theoretical underdevelopment. This is the case regardless of what the reasons are for this lack of (meta)theoretical aptitude. If we accept that a "theory is a scientific proposition [...] that introduces causal relations with the aim of describing, explaining, and/or predicting a set of phenomena" (Guest & Martin, 2021, p. 794), then the field-level theory that much of CCN work is based on has logical inconsistencies, namely manifesting as the formal fallacy of affirming the consequent. The ability to critically evaluate this was granted to us in part by the metatheoretical calculus, the formal model of the discourse, we built herein.

Based on our analyses, we propose that CCN as a subfield needs to reevaluate itself and take heed of calls for "cross-field fertilization" (Barak, 2017), especially in terms of theory. Our unique perspective here is to underline and explain why metatheorizing, specifically in the domain of formal logic, is just as important to consider — if not more so than methodological, computational, and theoretical issues in CCN since risking committing a formal fallacy in how we interpret our results is destructive to the whole enterprise. Just because a model correlates with neural and behavioral data, it is not sufficient for us to infer that the model is performing cognition: correlation does not imply cognition. We, (meta)scientists who contribute to CCN, must rethink how we reason about our work — by looking inwards, to understand how to move the subfield into a coherent state, and outwards, to learn how to perform metascientific reasoning from other established fields, like the super-field of cognitive science or the adjacent subfield of mathematical psychology.

If we do not examine the overarching principles that govern our science, we are ignoring the missing pieces to the puzzle (or indeed pizza; see figure 1, Guest & Martin, 2021) of why the field itself might not progress in intended ways. The goals of CCN involve looking at the lower-level mechanisms that give rise to neuroscientific findings and intelligent agents' behaviors (cf. Kietzmann et al., 2019; Shiffrin et al., 2020). If we allow flawed inference rules to govern CCN in a way that overlooks mechanism by ignoring the principle of multiple realizability (or if we permit the use of rules that are indeed formally fallacious), then we lead ourselves astray. To avoid this, we must as a field explicitly engage with the principle of multiple realizability, with theory building, and with the metatheoretical inferences we draw based

on our work, especially modelling work. The theoretically important stages of modelling work must not be forgotten, especially within the connectionist paradigm, and involve "exploring the effects of [experimental manipulations] on the model's behavior, and finally extracting implications of the simulation work for cognitive-level theory." (Guest et al., 2020, p. 290). However, it remains to be seen if indeed "ANNs and [biological neural networks] belong to the same family of direct-fit models" (i.e., models that use brute-force optimization to map input to output, Hasson et al., 2019, p. 417) and how such comparisons contribute to our understanding of cognition. Having a good grasp of the formal(izable) rules we use to reason, which are shaped by and in turn shape the ways in which we think about our science, will lead to a better understanding of cognition and the brain mechanisms that realize it — the core goals of the cognitive and neuro-sciences.

Acknowledgements The authors thank Ashley G. Lewis and Esther Mondragón for comments on an earlier version of this work. The authors are very grateful for discussions with and feedback provided by Sam H. Forbes, Iris van Rooij, Britta U. Westner, and Andy J. Wills.

Funding Olivia Guest and Andrea E. Martin were supported by the Netherlands Organization for Scientific Research (grant 016.Vidi.188.029 to AEM). AEM was supported by a Max Planck Research Group and a Lise Meitner Group "Language and Computation in Neural Systems" from the Max Planck Society.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, *46*, 1–6.

Blokpoel, M., Wareham, H., Haselager, W., Toni, I., & van Rooij, I. (2018). Deep analogical inference as the origin of hypotheses. *The Journal of Problem Solving*.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., & Biscione, V. (2022). Deep problems with neural network models of human vision.

Broadbent, D. (1985). A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General*.

Chalmers, D. J. (2020). What is conceptual engineering and what should it be? *Inquiry*, 1–18.

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, *199*(1), 767–790.

Chollet, F. et al. (2015). *Keras*. https://keras.io.

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

Collins, P. J., & Hahn, U. (2020). We might be wrong, but we think that hedging doesn't protect your reputation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(7), 1328–1348.

Collins, P. J., Krzyżanowska, K., Hartmann, S., Wheeler, G., & Hahn, U. (2020). Conditionals and testimony. *Cognitive Psychology*, *122*, 101329.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? on the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*.

Cummins, R. (2000). In F. Keil, & R. Wilson (Eds.) *Explanation and cognition*, (pp. 117–145). Cambridge: MIT Press.

Devereux, B. J., Clarke, A., & Tyler, L.K. (2018). Integrated deep visual and semantic attractor neural networks predict FMRI pattern-information along the ventral object processing pathway. *Scientific Reports, 8*(1).

Dujmović, M., Bowers, J., Adolfi, F., & Malhotra, G. (2022). The pitfalls of measuring representational similarity using representational similarity analysis. *bioRxiv*.

Dujmović, M., Malhotra, G., & Bowers, J. (2020). What do adversarial images tell us about human vision? *bioRxiv*.

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & et al. (2018). Adversarial examples that fool both computer vision and time-limited humans. arXiv:1802.08195.

Falkenburg, B., & Schiemann, G. (2019). *Mechanistic explanations in physics and beyond*. Berlin: Springer.

Firestone, C. (2020). Performance vs. competence in human-machine comparisons. *Proceedings of the National Academy of Sciences*, *117*(43), 26562–26571.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Frankfurt, H. G. (1958). Peirce's notion of abduction. *The Journal of Philosophy*, *55*(14), 593–597.

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S., & Bethge, M. (2020). The notorious difficulty of comparing human and machine perception. arXiv:2004.09406.

Geirhos, R., Meding, K., & Wichmann, F.A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. arXiv:2006.16736.

van Gerven, M., & Bohte, S. (2017). Editorial: Artificial neural networks as models of neural information processing. *Frontiers in Computational Neuroscience, 11*.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., & et al. (2021). Thinking ahead: spontaneous next word predictions in context as a keystone of language in humans and machines. *bioRxiv*.

Guest, O., Caso, A., & Cooper, R.P. (2020). On simulating neural damage in connectionist networks. *Computational Brain & Behavior*, *3*(3), 289–321.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *0*(0), 1745691620970585. (PMID: 33482070).

Harding, S. (1975). *Can theories be refuted?: Essays on the Duhem-Quine thesis* Vol. 81. Berlin: Springer Science & Business Media.
Hasson, U., Nastase, S. A., & Goldstein, A. (2019). Direct-fit to nature: an evolutionary perspective on biological (and artificial) neural networks.
Jason, G. (1988). Hedging as a fallacy of language. *Informal Logic, 10*(3).
Jonas, E., & Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLoS Computational Biology*, *13*(1), e1005268.
Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*(3), 339–373.
Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627.
Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, *180*, 101–109.
Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.
Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.
Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*.
Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.
Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160.
Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., & et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.
Lampson, B. (2006). Lazy and speculative execution. *Microsoft Research OPODIS, Bordeaux, France*, 21.
Leeds, D. D., Seibert, D. A., Pyles, J. A., & Tarr, M.J. (2013). Comparing visual representations across human fMRI and computational vision. *Journal of Vision*, *13*(13), 25–25.
Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife, 7*.
Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. arXiv:1807.06882.
Love, B. C. (2021). Levels of biological plausibility. *Philosophical Transactions of the Royal Society B*, *376*(1815), 20190632.
Luo, X., Roads, B. D., & Love, B.C. (2021). The costs and benefits of goal-directed attention in deep convolutional neural networks. *Comput Brain Behav*, *4*, 213–230. https://doi.org/10.1007/s42113-021-00098-y.
Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: towards using deep nets as models for human behavior. arXiv:2005.02181.
Mareschal, D., & French, R. M. (2017). TRACX2: a connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160057.
Martin, A. E., & Doumas, L. A. A. (2019). Predicate learning in neural systems: using oscillations to discover latent structure. *Current Opinion in Behavioral Sciences*, *29*, 77–83.
Martin, A. E., & Doumas, L. A. A. (2020). Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190306.
Marx, K. (1894). *Capital: volume III*. International Publishers, NY.
Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language*, *27*(2), 213–234.
McClelland, J., & Botvinick, M. (2020). Deep learning: Implications for human learning and memory. *PsyArXiv*.
Meijer, G. (2021). Neurons in the mouse brain correlate with cryptocurrency price: a cautionary tale.
Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*, 174569162097476.
Nicholson, D. A., & Prinz, A. A. (2020). Deep neural network models of object recognition exhibit human-like limitations when performing visual search tasks.
Nietzsche, F. (1886). Beyond good and evil. In (chap. Chapter IV (Apophthegms and Interludes)). Friedrich Nietzsche Internet Archive (marxists.org).
Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*(4), 647–670.
Peterson, J. C., Abbott, J. T., & Griffiths, T.L. (2016). Adapting deep network features to capture psychological representations. arXiv:1608.02164.
Plutynski, A. (2011). Four problems of abduction: A brief history. *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, *1*(2), 227–248.
Potochnik, A., & Sanches de Oliveira, G. (2019). Patterns in cognitive phenomena and pluralism of explanatory styles. *Topics in Cognitive Science*, *12*(4), 1306–1320.
Putnam, H. (1967). Psychological predicates. *Art, Mind, and Religion*, *1*, 37–48.
Quilty-Dunn, J. (2020). Polysemy and thought: Toward a generative theory of concepts. *Mind & Language*, *36*(1), 158–185. https://doi.org/10.1111/mila.12328.
Quine, W. V. (1951). Main trends in recent philosophy: Two dogmas of empiricism. *The Philosophical Review*, *60*(1), 20–43.
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.
Reverberi, C., Pischedda, D., Burigo, M., & Cherubini, P. (2012). Deduction without awareness. *Acta Psychologica*, *139*(1), 244–253. https://doi.org/10.1016/j.actpsy.2011.09.011.
Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). How hard is cognitive science? In *Proceedings of the annual meeting of the cognitive science society*.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, *107*(2), 358–367.
Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & et al. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205.
Russell, B. (1918). *The philosophy of logical atomism*. Evanston: Routledge.
Salmon, M. H. (2013). *Introduction to logic and critical thinking* (6th ed). Cengage Learning.
Saxe, A., Nelli, S., & Summerfield, C. (2020). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 1–13.
Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M.A. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, *181*, 775–785.
Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1*(1), 1–17.
Shiffrin, R. M., Bassett, D. S., Kriegeskorte, N., & Tenenbaum, J.B. (2020). The brain produces mind by modeling. *Proceedings of the National Academy of Sciences*, *117*(47), 29299–29301.

Sundholm, G. (1994). Existence, proof and truth-making: A perspective on the intuitionistic conception of truth. *Topoi*, *13*(2), 117–126.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & et al. (2014). Intriguing properties of neural networks.

Tacchetti, A., Isik, L., & Poggio, T. (2017). Invariant recognition drives neural representations of action sequences. *PLoS Computational Biology*, *13*(12), e1005859.

Truzzi, A., & Cusack, R. (2020). Understanding CNNs as a model of the inferior temporal cortex: using mediation analysis to unpack the contribution of perceptual and semantic features in random and trained networks.

Turing, A. M. (1950). Computing machinery and intelligence. *Creative Computing*, *6*(1), 44–53.

Vickers, P. (2019). Towards a realistic success-to-truth inference for scientific realism. *Synthese*, *196*(2), 571–585.

Wald, H. (1975). *Introduction to dialectical logic* Vol. 14. John Benjamins Publishing.

Wray, K. B. (2013). Success and truth in the realism/anti-realism debate. *Synthese*, *190*(9), 1719–1729.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. arXiv:1611.03530.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*(1), 1–9.

Zipser, D., & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*(6158), 679–684.