Scholarly AI system diagrams as an access point to mental models

Guy Clarke Marshall¹, Caroline Jay¹, and André Freitas¹,²

¹ Department of Computer Science, University of Manchester, Manchester, UK
guy.marshall@postgrad.manchester.ac.uk,
{caroline.jay, andre.freitas}@manchester.ac.uk
² Idiap Research Institute
Rue Marconi 19, Martigny, 1920, Switzerland

Abstract. Complex systems, such as Artificial Intelligence (AI) systems, are comprised of many interrelated components. In order to represent these systems, demonstrating the relations between components is essential. Perhaps because of this, diagrams, as “icons of relation”, are a prevalent medium for signifying complex systems. Diagrams used to communicate AI system architectures are currently extremely varied. The diversity in diagrammatic conceptual modelling choices provides an opportunity to gain insight into the aspects which are being prioritised for communication. In this philosophical exploration of AI systems diagrams, we integrate theories of conceptual models, communication theory, and semiotics. We discuss consequences of standardised diagrammatic languages for AI systems, concluding that while we expect engineers implementing systems to benefit from standards, researchers would have a larger benefit from guidelines.

Keywords: Neural Networks · Systems · Diagrams · Conceptual Models

1 Introduction

Diagrams are a signifier, cognitive aid, and communication channel. In describing software systems, diagrams often provide a level of abstraction that facilitates an understanding of the overall structure, and the relation between the computational artifacts of the system. Software system diagrams have a dual role bridging between cognition and communication of humans, and representation of mechanisms entailed by machines. In the words of Horn [10]: “Visual language has the potential for increasing human bandwidth, the capacity to take in, comprehend, and more efficiently synthesize large amounts of new information”.

In the field of AI, the pace of development is high, and as such conferences are the most prestigious academic venues. In their scholarly proceedings, we find that the majority of papers include a system architecture diagram by way of structural explanation. This is common across Computer Science and other system-centric domains. What is less common is the variety of these representations, even when compared with other research and engineering domains. Despite being based
on similar and mathematically-well-formalised computational artifacts, such as
neural networks, the diagrams have very low consistency. In this paper, we utilise
the opportunity provided by the lack of convention to gain insight into the way
the creators of AI systems are choosing to communicate their systems.

This study is motivated by a number of questions. Why are diagrams being
used to describe NN systems? Why are the diagrams so heterogeneous? What
would be the impact if this heterogeneity was reduced through standards or
guidelines? In order to make progress towards these broad questions, we inves-
tigate the following hypotheses:

- H1: There is a relation between content included in NN system diagrams
  and their role in the scholarly community as conceptual models.
- H2: There are visual encoding prioritisations which align with subcategories
  within mental models theories.
- H3: At present, guidelines for diagram creation would be beneficial for com-
munication.

2 Background

AI systems are software, written in a programming language, and often based
on neural networks. A neural network takes an input (usually text, numbers,
images or video), and then processes this through a series of layers, to create an
output (usually classification or prediction). Each layer contains a set of nodes
which hold information and transmit signals to nodes in other layers. Specific
mathematical functions or operations are also used in these systems, such as
sigmoid, concatenate, softmax, max pooling, and loss. Different architectures
are used for different types of activities: Convolutional Neural Networks (CNN),
inspired by the human visual system, are commonly used for processing images,
and Long Short Term Memory networks (LSTM), a type of Recurrent Neural
Network (RNN) which are designed for processing sequences, are often used for
text. These neural networks “learn” a function, but have to be trained to do so.
Training consists of providing inputs and expected outputs, so the system can
learn how an input should be processed. The system is then tested with unseen
inputs, to see if it is able to process these correctly. The data perspective (focused
on vectors and matrices) and the functional perspective are fairly distinct ways of
thinking about what is happening, which is part of the problem of communicating
in this area. The architectural perspective on the system can be to a greater
or lesser extent encompassing the data manipulations, which leads to a broad
spectrum of possible representations.

In terms of communicating about AI systems, it is not just from researchers
to the public, but also between researchers that attention is needed. The act of
communicating between researchers is not only about making code reproducible
[11], though we agree that is important, but making cognitively accessible the
systems and the computational artifacts of which they are comprised. As a result
of this, diagrams have a role to play in the ethics of AI, and are a key way of
establishing transparency and managing risks around what the system is doing.
3 H1: There is a relation between content included in NN system diagrams and their role in the scholarly community as conceptual models

To discuss this hypothesis, we first describe the heterogeneity in representation, and then proceed to classify the observed phenomena, linking the observed diagrams to conceptual and mental models.

3.1 Heterogeneity in representation

In principle, the primary purpose of diagrams in scholarly publications is communicative. The authors are attempting to communicate through a diagrammatic medium some kind of relational structure. Diagrams are ideally suited to this task, and are used for this purpose in many domains [17]. However, there are other social aspects. Without passing too critical an eye over the scientific endeavour, having a “good looking” diagram (both aesthetically and technically) may improve perception of a paper, thereby making it more likely to pass through peer review. While the visual encoding methods are quite unconstrained and heterogeneous, it is conventional to include an architecture diagram if the paper presents a new system. The subsequent discussion attempts to make steps towards understanding the reasons for, and consequences of, this heterogeneity.

Perhaps due to the complexity of these systems, there are a number of diagrammatic representational choices that are made by different authors attempting to express different things. Given the social nature of research, it is curious that there is not more prevalent “copying” or adoption of informal diagrammatic encoding conventions, or even convenience-based similarity caused by the use of popular diagramming tools. Given that it would be practically easier for authors to directly copy existing styles, it is unlikely to be chance, but rather we argue that there must be a compelling reason for authors to be creating such different diagrammatic representations.

A partial explanation for heterogeneity could be a lack of appropriate diagramming tools. In a recent interview study involving technical domain experts, Ma’ayan et al. [17] found that “To illustrate concepts effectively, experts find appropriate visual representations and translate concepts into concrete shapes. This translation step is not supported explicitly by current diagramming tools”. This does not explain why AI software system diagrams are so heterogeneous compared to other domains, nor does it explain the lack of informal conventions.

Nefdt [19] argues that the state of understanding of AI systems is such that “meaningful components” and compositional structures have not been established for deep neural networks. He does not extend his argument to the entire system, focusing instead on the mathematical “black box” that is a deep neural network. If we adopt this stance over an increased scope, this would seem to support the claim that there is an underlying cognitive reason for the differences in representations. As such, we argue that insight into the current state of understanding can be gained partially as a result of the current “epistemic opacity”, and therefore allows an insight into the representational priorities of the author.
We hypothesise that a particular representational aspect is prioritised by the author either because it shows what they think is important, or because it is what they would want to see in a diagram authored by their peers. In either case the priority is effective communication. Differences in prioritisation may be causing the creation of bespoke diagrammatic encodings. When representing an AI system, the diagram author can prioritise different aspects of what is represented:

- Function: Operations which occur, representation transformations, and the purpose of parts of the system
- Data: The data model, type, dimensionality and how it is manipulated
- Example: Showing the steps of an example input through that system
- Contribution: Focusing on the scientific novelty of the approach, giving much more detail in that area
- Code: Important class names and the order in which they are called
- Mathematics: Including specific mathematical functions
- Index to text: Using a label structure to allow for easier referencing

These representational priorities result in different content being displayed through different visual encoding mechanisms. It is usual to have aspects of several of these priorities, as it is not the case that the prioritisation of one aspect necessarily inhibits another. In terms of how the diagrams are presented within a paper, some papers include multiple diagrams, either by multiple figures or sub-figures. Often, sub-figures or boxes are used to give both schematic and detailed views within the same diagram. Dependencies are often indicated by arrows. Diagrams almost always represent important content in natural language, such as labels or descriptions.

3.2 The content of diagrams

The inclusion of an “example” makes the diagram be “of” a particular run of that system. However, it is understood that this is signifying the system itself. There are some diagram users who take an example instantiation and use this to generalise to the operation of the overall system [18]. Indeed, in this way the diagram is supporting logical induction, rather than deduction from general rules (which would be more classically descriptive of Function). Similarly the omission of inputs and outputs (e.g. “text” or “probability”) from the diagram makes the diagram be “of” a set of operators rather than of a functioning system, which is equally understood as signifying the system (indeed, including data).

Rarely are AI systems’ Knowledge Representations (links with semantics) represented visually in this scholarly community. Instead, labels are favoured for this important system component. In some diagrams, they include the dimensions of the embedding space, a drawing of an arbitrary graph, or a simplified example of part of the knowledge representation. The high dimensionality and the sparsity involved makes visual representation challenging.
3.3 Specific graphical objects

Arrows have been linked with the concept of functional processes \[8\]. All surveyed diagrams used arrows, perhaps suggesting that the systems are understood as being compositional and sequential (rather than as objects which enact a function).

Labels for “layers”, sets of neural nodes often performing a particular neural network function, are common. These neural network functions often have metaphorical or descriptive names, such as “embedding” or “attention”, which form an informal lexicon of computational artefacts. The labels are often used alongside more complex sets of graphical objects signifying the system, performing the role of (linguistically) simplifying and linearising the system without taking a sentential form. To take a specific example, the layer labels are often of the form “embedding layer”, i.e. mentioning the function. This in turn might be labelling a set of blocks labelled “BERT”. In this sense, this type of diagram contains cognitive support for two parallel mental models, that of function in the original diagram, and that of state in the layer labels. This particular aggregation-and-switching-mental-model usage of labels is very common, but not exclusive. Some diagrams may instead have mathematical functions (e.g. “tanh”) labelling layers. Principally the association between these representations is grouping by alignment, but may additionally make use of “linking” graphical objects such as arrows, brackets, or blocks, or grouping by colour.

Many graphical objects, and almost all diagrams, utilise labels (in English). This suggests a low level of semantic content within the graphical objects. Perhaps this is because of the non-physical nature of the systems, which makes it challenging to signify visually without the aid of an established formal language.

4 H2: There are visual encoding prioritisations which align with subcategories within mental models theories

This subsection aims to make progress towards the classification of AI system diagrams. The link between diagramming priorities and mental models is speculative, but could serve to explain some of the manifest heterogeneity.

4.1 Mental models and conceptual models

We can further refine this schema to correspond more directly to a theory of mental models. Definitions of mental models are various, but can perhaps be summarised by: “A mental model is a simplified representation of reality that allows people to interact with the world.” \[14\]

In Computer Science, Guarino et al. \[7\] recently developed a model for the relationship between mental models and conceptual models. For Guarino et al., mental models are “personal, partial accounts of the external reality, filtered through the lens of a conceptualization, that people use to interact with the world around them.” A mental model (or perhaps less ambiguously a conceptual mental representation), is described by a conceptual model, in our case the
system diagram with a communicative purpose. “Note that, being an information object, a conceptual model is always the result of an intentional act. In other words, conceptual models are artifacts produced with the deliberate intention of describing a conceptualized reality.” With further specific reference to the Computer Science domain, Guarino et al. state that “we may say that a computer program is a conceptual model of the computer’s internal behavior, but only as long as its programming language’s primitives denote concepts concerning computer behavior. If they rather denote data, we conclude that such a computer program is not a conceptual model.” Guarino et al.’s work highlights the challenges of philosophical precision in computer science when discussing mental models. Applying these definitions to AI systems diagrams, it is variable whether the diagram is a conceptual model. As such, in this work we use mental models as a metaphor, rather than in the strict sense. Usage of the term in this manner is common practice in human-computer interaction.

To summarise our working definitions: The conceptual model is the artefact itself, with communicative purpose, to articulate concepts and the relationships between concepts. Diagrams are often used for this purpose, and have been found to aid learning and reasoning [17]. In our usage, a mental model is the way an author is thinking of the system when creating that conceptual diagram for use in scholarly publication. We draw on psychological mental model research, itself utilising a more rigorous definition, in order to assist in the classification of conceptual diagrams, because the fit appears natural. We explore potential consequences if this is aligned. However, we do not suggest a deep claim about mental operation or cognition, which would require an empirical study.

4.2 Classification using mental models theory

Fig. 1. Representational choices in diagrams of AI systems
In this work, we are focusing on the diagrammatic metaphor being used for what is represented rather than the visual encoding. Bringing types of AI software system diagrams closer to mental models, we can draw parallels with mental model categories \cite{22} and types of diagram observed. Diagrams may have aspects of multiple categories.

- **Function**: Explaining how the system operates by emphasising functional aspects, such as mapping, input and output. Operations used as a verb. For example, “word embedding” (a general term) rather than “BERT” (a specific architecture for embedding \cite{6}). This type often includes example input and output.

- **Contribution (Purpose)**: Omits the majority of information other than that required to understand the sub-section of the system that contains the novelty of the system or approach.

- **Schematic (Form)**: Describes the system at a high level, uses probably a block-style without iconic graphical objects. Does not include mathematical or data details. The schematic may relate to classes or packages used. In order to be distinct from Function, it commonly uses e.g. “BERT” rather than “word embedding”. Also often aggregates graphical objects into modules. (In order to be “state” (i.e. what it is doing) the block diagram should be verbs).

- **Data manipulation (State)**: Includes data dimensions, and usually a visual representation of the data itself. It describes what the system does to the data, so this also includes where operations are primarily labelled arrows (rather than inside blocks).

- **Example**: How example data transforms. Includes example input, output and intermediate steps. Usually better relates to Function (how it operates) rather than Form (what it looks like), but this depends on the graphical objects used. It is useful to disambiguate this, particularly for Image Processing, where often the diagram is a visual representation of the data manipulation. In one sense this is the Form of the example data manipulation, and in another sense the Function of the system on a single example. It is the latter that we are concerned with, in our assertion that the diagrammatic representation is signifying the system rather than the example. Note also that the inclusion of intermediate step as using the same example is important. Many diagrams include example inputs and outputs, but in the processing of the system they do not utilise the example and the diagram can be “schematic”.

Sometimes the diagrammatic representations found in conference proceedings contain errors. In addition to typographic errors, these can be visualisation errors, in the sense that the diagrams may cause confusion or inaccurately reflect the reality of the underlying system. For example, the circles representing vectors can represent a precise number of objects, or not. Fig. \ref{fig:example} shows an example where the pairs of circles represent two LSTM output vectors (a common representational choice), while the three circles of $\hat{P}$ do not represent three vectors but rather $j$ vectors, where $j$ is the number of words in the sentence. The
omission of the ellipsis following the embedding layer appears to have led to this visual encoding choice. This unmeaningful 3-vector is repeated perhaps more dangerously in the final concatenation before “multi-feedforward”. This can be understood by careful reading of the words and formulae in the text, but could be misleading, as found in an interview study by Marshall et al. [18].

Fig. 2. An example diagram, used by Li et al. [16], using labels for Form, graphical objects variously for Form and State, input and output for Purpose.

We claim these potential different forms of diagrammatic representamen are signifying the same referent, a software system, through different lenses. We are not claiming that precisely the same sense is made with different diagrams. Even strictly isomorphic representations have different semiotic properties, such as the perceptual and cognitive properties of mathematics as performed by equation, Euler diagram or language, or indeed without the ability to perform gestures.

The separation of Form from Function is neither self-evident nor entirely natural. In some software paradigms, such as functional programming, the essence of Form and Function are entangled. By being faithful to the perspective of the system’s Form and Function, rather than to the human cognitive process which it is automating, we are able to be crisper with this distinction.

The types of representamen employed are of particular interest. Our hypothesis is that the diagrams, and the variety we see exhibiting the above principalities, are a result of the range of cognitive functions being employed by different users. At present, there is no common language they are using to communicate. No suitable representation providing as cognitive support such as symbolic equations gave to the mathematicians of ancient Babylon, or the letter x gave to Descartes [5], has been sought nor found. In Physics, Coecke and Kissinger [4] use diagrams to reason about quantum processes and to diagrammatically perform the calculations required to understand them.

Another important aspect is reproducibility. From diagrams, the building of great works of industrial-revolution engineering such as the SS Great Britain was possible [24]. Indeed, the technique to perform diagrammatic projection from 3D to 2D is seen as one of Marc Isambard Brunel’s main contributions [20]. Circuit diagrams and other standard diagrammatic representations, often implemented
or overseen by professional bodies, have also enabled this standard form of communication and reproducibility across many domains, including Computer Science. With engineering diagramming technology in mind, we proceed to consider specific graphical objects used in AI system diagrams.

4.3 Mental operations

Ryan Tweney’s lifetime of work examining Faraday’s cognitive processes through images provides insight into a single author’s cognitive process (see Ippolito and Tweney [12]). Without the breadth and depth of work of a single author, it is not practical to carry out a similar analysis in AI systems. Nevertheless this work served to inspire the approach taken here, and necessitates brief discussion of “mental models” in the cognitive science sense.

All three primary mental operations of apprehension, judgement, and inference [9] are at play in using AI system diagrams for research. Apprehension, being the forming of a picture in one’s mind, is important for depicting and understanding of the system. Any example of Judgement could be “this is relevant”, and Inference “this is a contribution”. We focus on the creation of mental representation, i.e. Apprehension. Note that “Mental models can be constructed from perception, imagination, or the comprehension of discourse” [13], and conference proceeding AI software system diagrams are in some sense all three of these: Perception of the diagram, imagination of the system operation, and comprehending scientific discourse.

We can categorise mental models as being about Purpose, Function, State and Form [21], which has been directly applied to complex systems in a team context [22]. In this sense,

– Purpose is why a system exists;
– Function is how the system operates;
– State is what the system is doing;
– and Form is what the system looks like.

These categories of mental models can be related to the types of diagrams found in AI systems literature, as shown in Fig. [1] As highlighted by Rouse and Morris [22], these facilitate describing, explaining and predicting: “Mental models are the mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions (or expectations) of future system states.”

At first glance, it may seem that the correspondence between mental models is not strong for the “Function” model. As noted previously, in what is observed some diagrams prioritise an instantiated example. This can be interpreted as a different way of reasoning (inductive rather than deductive), and in both cases appears to be focused around articulating Function. The detail-centric vs function-centric comparison partially describes how abstract or schematic the diagram is. A difference might be the inclusion of a block labelled as “CNN”, only including hyperparameters in the former. The inclusion of hyperparameters,
which are non-structural configuration details, demonstrates that the diagram supports a different level of reasoning.

Aspects related to mental models:

- **Purpose** is why a system exists, specifically the activity it performs. There are canonical tasks often linked to human cognitive operations (such as Named Entity Recognition or Image Classification). The discipline is evolvable and systems are comparable because there are these canonical tasks. **Manifestation:** Often evidenced by inputs and outputs. Title, caption, other language or images describing the task.

- **Function** is how the system operates. **Manifestation:** The inclusion of system operations as category terms. Graphical objects representing nodes. Mathematical equations.

- **State** is what the system is doing, with temporal elements. In AI systems, this is about data states at a particular functional position. **Manifestation:** Often shows how the dimensions of the data are changed, and usually a visual representation of the data itself. It describes what the system does to the data, so this also includes where operations are primarily labelled arrows (rather than inside blocks).

- **Form** is what the system looks like, and how the system is arranged. It is succinct, expressive and general, comprised of the representational and functional choices, how they are structured and composed, with emphasis on the dependencies between components. The physical structure analogy is perhaps closest to the classes and modules of code, and the shape of the data. This is usually succinct. **Manifestation:** The inclusion of system components as nouns. Often aggregating into labelled modules.

Note that a State Diagram in Computer Science is used to describe a system with a finite number of states and their transitions, is available in a standardised form in UML [2]. Whilst theoretically this could be used to describe many AI systems, empirically this is a very uncommon scholarly choice of representation.

### 4.4 Abstraction levels

![Abstraction levels](image)

**Fig. 3.** Abstraction levels used in interacting with AI systems
Fig. 3 shows part of the context of abstractions and formalisms surrounding system architecture diagrams, particularly in the absence of a formalism to describe functional modules, or specified visual encoding frameworks. It is quite possible that the system creator does not have a detailed knowledge of how their code will execute (or even where, in the case of cloud computing). Of course, mathematical formalisms and related work may also be expected to form part of the author’s model of the system. With the complexity of AI software systems, it is unlikely to be practical to perceive both the granular underlying mathematics and the overall system structure in one instant. Indeed, the structural explanation given by a diagram is necessarily schematic, omitting information in order to be more efficient. In Fig. 3, each of the arrows has some loss. Not all aspects of the author’s model are included in the system architecture diagram, for example.

In a single diagram, there can be different levels of granularity. Graphical representations of vectors are employed alongside higher level routines such as biGRU or LSTM, themselves often comprised of multiple vector operations. The different abstractions and granularity levels available to diagram authors facilitates the diversity of representations that are observed, both within a single diagram and across the diagrammatic corpus.

4.5 Diagram formats representing the same system

![Diagram](image)

Fig. 4. Example representations of the same system, obfuscating detail and graphical encoding differences
Fig. 4 shows examples of the overall structure of diagrams commonly found in AI systems. These schematics omit many details and complications of “real” diagrams. See Fig. 2 for a typical example from ACL 2018, noting the lack of linearity and variety of visual encodings used even within a single diagram. Whilst the two versions of Function and State represent equivalent content (in that there exists an isomorphism to convert between them), they are not identical. Each structure places emphasis on different aspects of the representation.

We can apply this framework to real examples. Fig. 2 shows a diagram taken from proceedings of a top conference, describing the system architecture that is their contribution. The diagram has labelled layers describing Form, limited Purpose (example input and output, with no intermediate steps, but the labels of “plot” and “ending” facilitate abduction), and the majority of diagrammatic real estate given over to graphical elements describing vectors (the circles). Independently of the text, abbreviations such as $s_i$ may be understood by their context and abduction. Concatenation is a (nominalised) verb on arrows, while Feature Extractor, element-wise product, and the majority of other labels are nouns. Combined with the emphasis on the data/vectors themselves, although they do not have dimensions indicated either numerically or pictorially, it would seem this diagram is primarily State-based, with Form labels. We hypothesise the author is thinking about the system as manipulating data, as they prioritise communicating State but without much information about the data itself (just grouping by colour, and repetition for scale). This suggests that the author feels the “important” part of the system to communicate is not necessarily the way the data changes but the overall system. We hypothesise this “perceived importance” is in itself a window into the mental representation of the authors: In the absence of any independent advice into effective diagrams, authors using “examples” to instantiate their cognition about a system will see it as important for others to include “examples” in their own diagrams.

5 H3: At present, guidelines for diagram creation would be beneficial for communication

Hypothesis H3 arose from the prior analysis, combined with reasoning concerning the consequences of a formal language or standards for AI system diagrams, balancing this against benefits associated with some level of convention or familiarity, such as the reduction of “Semiotic-Semantic failure in encoding” [15].

There are two aspects relating to the utility of standards or conventions. The first is uptake, the second is effectiveness. In the domain of Human Computer Interaction (HCI), where standards are commonplace, this would be seen as usability and utility. In HCI, the “standards versus guidelines” distinction is known, with the flexibility of guidelines being important for software [23]. A more granular and up-to-date comparison between design guidelines, style guides, principles and patterns is summarised by Al-Sa’di [1], with a breadth of citations.
If standards were enforced, for example by publishers or conference organisers (such as ACL, the Association for Computational Linguistics) then the concern would be whether those standards genuinely improved the quality of the communication. Standards would facilitate comparison between systems, facilitating inter-diagram reasoning. If thoughtfully composed, standards could be expected to streamline the diagram creation and interpretation processes. However, by using standards, we would risk losing the author-specific window into what is specifically important about the diagram in question. For particularly novel architectures that make use of some hitherto un-thought-of function, there are additional challenges, as the existing standards may not facilitate effective representation of this novel element. Further, we risk authors not being able to communicate their mental model. However, it may be that readers have different ways of thinking about the systems, and that a different diagram prioritising communication of different aspects would be more pragmatically effective.

For well established architectures, standards would restrict the ability of talented diagram authors to innovate. AI systems are rapidly developing [3]. The ability to create novel representations in a rapidly changing domain such as AI systems is of value, as the representational requirements are also changing. The useful information about an architecture can even change. Initially, it may be that the key thing is to communicate novelty and utility, primarily by getting the surrounding paper accepted into a prestigious academic venue. In later stages of the system’s communicative “life”, the key thing to communicate may be how to use, build, or adapt the system, in order to obtain a higher number of citations and higher impact.

For the reader, standards or conventions would help avoid the barrier of interpreting each individual author’s way of representing things. They would reduce ambiguity, and reduce the potential for misunderstanding. In this way, we would expect standards to make the diagram interpretation significantly more efficient. Similarly, diagram creation would be more straightforward for authors, as the content, graphical objects and method of arranging them becomes more prescribed. The challenge is whether the content prescribed as being included in the diagram represents what is useful to be communicated about the system in context of the wider communicative system.

In order to maximise the usefulness of any standards or conventions, it would be necessary to understand deeply the requirements of the community, for authors and readers, and to have an adaptive mechanism for updating the standards as the requirements of the community change. We might expect this to happen naturally, not being consciously or scientifically based on the needs of users, but rather being optimised for authors and the peer review process.

For those involved in engineering of systems, the arguments in favour of standard diagrammatic languages are strong. The allowance for comparison, scale, precision, unambiguity, completeness and replicability, such as is available to Mechanical Engineers, for example. It would be necessary that the standards would allow communication about novel systems while complying with those standards.
For researchers primarily concerned with a higher level of abstraction (rather than system-building), the arguments are more nuanced. In their heterogeneity, there is a window into the communicative goal of the author, allowing the readers to gain some insight into their conception of what their system is, how it works, and the contribution it makes. This insight is crucial for scholars in order to deeply understand the publication. It would be theoretically possible for representation to be facilitated by a domain-specific language, as is done in mathematics or in quantum mechanics, as described in Section 4. For AI systems, due to the prevalence of relations between components involved, it would be intuitive to have this language be diagrammatic. In the absence of such a mature language, there is still the important issue of having to understand each individual author’s way of diagrammatically encoding their system. Comparison between systems, important for positioning the contribution of the new architecture against related work, is difficult. Guidelines, a set of principles providing direction for designing diagrams, would aid in reducing this cognitive barrier, and reduce misunderstandings, whilst still affording the ability to customise depending on the nature of the author’s representational requirement. Due to this, we advocate the creation of unenforced diagram guidelines.

Guideline adoption could be organic, and their creation involving the community, with inbuilt mechanisms for adaptation and evolution. This would ensure higher effectiveness and longevity, at the expense of slower uptake. The slower uptake would reduce effectiveness since, at least in the short term, the format would be unlikely to be widespread. A gradual movement towards consistency where appropriate, particularly if we are able to draw from the community of diagrams researchers a “fitness function” for assessing these guidelines, would appear to be the most effective solution. This approach would facilitate evolution, to create better representations that better support the representational requirements of the community. Indeed, it may be that future diagrammatic representation of systems will be interactive. Even in this case, guidelines could support the design of this new medium. Whether ultimately the solution is guidelines, standards, conventions, or a new representational technology, remains to be seen. For AI system engineering, when the field is at an appropriate level of maturity, a standardised language would likely be very useful. From a research perspective, the nature of AI system research is such that it is inherently changing, with new architectures being designed which can be expected to have novel representational requirements. Unless the field of AI system research stabilises in terms of architectural representational requirements or changes focus, thereby becoming something different to it is at present, a standardised language would be inappropriate. Instead, evidence-based guidelines should be encouraged.

6 Conclusion

We argue that the heterogeneity in diagrammatic representations of AI systems is due to the inherent complexity of what is being represented and a lack of obvious good representational choice for the system themselves. In the case of many other
scientific and engineering disciplines, a standard has quickly emerged. It may be that we are too early in the life of “AI science and engineering” to see this, and instead are able to gain insight from the heterogeneity. The heterogeneity seen today is a manifestation of a lack of conventional system elements and visual encoding principles.

AI systems are a new medium, which at its lowest representational level are not easily interpretable. Currently there is heterogeneity in diagrammatic representation. Diagrams are a useful, and efficient, way of understanding AI systems. Diagrams are a fundamental signification layer, encoding the design and metadata description of these emerging systems. We have linked this to a morphic semiotic perspective: “code interpretation” being more computationally expensive than “diagram interpretation”. We have hypothesised the diagrams used at present to have relationship to mental models. As the community evolves, and the representational requirements of these diagrams become clearer, as with many languages, we expect some consolidation around effective ways of signifying the underlying system, and providing cognitive support for communicating and reasoning about these systems.

Finally, we encourage the creation of evidence-based guidelines for AI system architecture diagrams, that are designed to be usable for creators of novel architectures, and created mindful of the needs of diagram users.

References

[1] Al-Sa’di, A.: User interface guidelines for Tablet PC Arabic educational applications. Ph.D. thesis, Auckland University of Technology (2018)
[2] Booch, G., Rumbaugh, J., Jacobson, I., Wesley, A.: Unified Modeling Language User Guide, The. Tech. rep., Rational Software Corporation (1998)
[3] Chen, M., Herrera, F., Hwang, K.: Cognitive computing: architecture, technologies and intelligent applications. IEEE Access 6, 19774–19783 (2018)
[4] Coecke, B., Kissinger, A.: Picturing quantum processes. Cambridge University Press (2017)
[5] Descartes, R.: Discours de la methode pour bien conduire sa raison, et chercher la verite dans les sciences. Plus la Dioptrie. Les Meteores. Et la Geometrie. – Qui sont des essais de cete Methode. De l’Imprimerie de Ian Maire (1637)
[6] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)
[7] Guarino, N., Giuzzardi, G., Mylopoulos, J.: On the philosophical foundations of conceptual models. Information Modelling and Knowledge Bases XXXI 321, 1 (2020)
[8] Heiser, J., Tversky, B.: Arrows in comprehending and producing mechanical diagrams. Cognitive science 30(3), 581–592 (2006)
[9] Hobhouse, L.T.: The Theory of Knowledge (Routledge Revivals): A Contribution to Some Problems of Logic and Metaphysics. Routledge (2013)
[10] Horn, R.E.: Visual language and converging technologies in the next 10-15 years (and beyond). Converging Technologies for Improving Human Performance 52(7), 124 (2002)
[11] Hutson, M.: Artificial intelligence faces reproducibility crisis. Science 359(6377), 725–726 (2018), ISSN 0036-8075, https://doi.org/10.1126/science.359.6377.725
[12] Ippolito, M.F., Tweney, R.D.: The inception of insight. In: The nature of insight, chap. 13, pp. 433–462, The MIT Press (1995)
[13] Johnson-Laird, P.N.: Mental models: Towards a cognitive science of language, inference, and consciousness. 6, Harvard University Press (1983)
[14] Jones, N.A., Ross, H., Lynam, T., Perez, P., Leitch, A.: Mental models: an interdisciplinary synthesis of theory and methods. Ecology and Society 16(1) (2011)
[15] Lanigan, R.L.: Information theories. Theories and models of communication 1, 59–83 (2013)
[16] Li, Q., Li, Z., Wei, J.M., Gu, Y., Jatowt, A., Yang, Z.: A multi-attention based neural network with external knowledge for story ending predicting task. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1754–1762 (2018)
[17] Ma’ayan, D., Ni, W., Ye, K., Kulkarni, C., Sunshine, J.: How domain experts create conceptual diagrams and implications for tool design. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2020)
[18] Marshall, G.C., Jay, C., Freitas, A.: Understanding scholarly natural language processing system diagrams through application of the richards-engelhardt framework. arXiv preprint arXiv:2008.11785 (2020)
[19] Nefdt, R.M.: A puzzle concerning compositionality in machines. Minds and Machines pp. 1–29 (2020)
[20] Rolt, L.T.C.: Isambard Kingdom Brunel: A Biography. Longmans, Green (1957)
[21] Rouse, W.B., Cannon-Bowers, J.A., Salas, E.: The role of mental models in team performance in complex systems. IEEE transactions on systems, man, and cybernetics 22(6), 1296–1308 (1992)
[22] Rouse, W.B., Morris, N.M.: On looking into the black box: Prospects and limits in the search for mental models. Psychological bulletin 100(3), 349 (1986)
[23] Smith, S.L.: Standards versus guidelines for designing user interface software. Behaviour & information technology 5(1), 47–61 (1986)
[24] Weale, J.: The Great Britain Atlantic Steam Ship. John Weale (1847)