Reflexivity in AI systems: metacognition, metalearning, self-improving systems, and second-guessing

Abstract: Our capacity for reflexivity enables us to introspect on our thoughts and preferences, engage in metacognition, and evaluate our progress on projects while reconfiguring our approach as needed. Despite being central to our human intelligence, it has received marginal attention in AI research. To be clear, AI systems exist that manifest reflexivity. As we argue, meta-reinforcement learning modules and self-improving systems manifest reflexivity, as do Large Language Models that have the capacity for metacognition, so-called self-attention, and second-guessing. Reflexivity is the common denominator in each of these capacities and mechanisms. The manifestation of reflexivity in current AI systems is, however, a side effect of other goals. To date, no explicit attention has been directed at artificially replicating this central element of human intelligence. We propose that a general capacity for reflexivity is key to bringing AI to the next level.

Keywords: Self-Attention Mechanisms, Metacognition, Meta-learning, Large Language Models, Self-Representation

Reflexivity is a central element of human intelligence. It allows us to have first-person thoughts and examine our reasons, preferences, and values. It enables us to introspect on our beliefs and emotions. Due to having the capacity for reflexivity, we can evaluate our plans and adjust them as needed. More generally, reflexivity allows an organism or system to represent (a part of) itself while registering that it is that organism or system. The capacity for reflexivity is key to bringing AI to the next level.

AI systems have already been developed here and there that manifest reflexivity. To illustrate, a system that explicitly includes itself in its model of the world manifests reflexivity, as does a Large Language Model that tracks its contributions to a conversation over time and can summarize or comment on them. While there is evidence of reflexivity in current AI systems, with few exceptions, it was generated as a side effect of aiming to develop other capacities, such as the capacity for a system to improve its architecture or safely navigate the environment autonomously. To date, no explicit attention has been directed at artificially replicating this central element of human intelligence.[1] Now, there have been calls for Al needing a prefrontal cortex (Russin et al., 2020; . The capacity for reflexivity is central to the functioning of the human prefrontal cortex. While we are far from generating an artificial prefrontal cortex, artificially replicating the capacity for reflexivity is within reach.

Here is the plan. Section 1 provides an analysis of reflexivity. Section 2 discusses ways to assess whether a system has the capacity for reflexivity. The rest of the paper zeroes in on manifestations of reflexivity in AI systems, specifically, metacognition, metalearning, self-improving systems, and second-guessing. Closely related systems that fail to manifest reflexivity are discussed to pinpoint what it takes for a system to manifest reflexivity rather than fail to do so.

But first, a note of caution. In humans, reflexivity may manifest in self-reference, self-representation, self-awareness, and first-person thought. Reflexivity is often discussed under those labels —along with the more technical term "*de se*" (García-Carpintero, 2024). To avoid any implication that possessing the capacity for reflexivity entails having a self, the term "reflexivity" is preferred over alternatives such as "self-representation," "self-reference," "self-awareness," or "first-person thought." After all, there are powerful reasons to deny that humans have a self, although we may have the illusion of having a self (Schellenberg, 2025). There are even stronger reasons to deny that an AI system has a self (Shanahan, 2024). The important point for the current discussion is that having the capacity for reflexivity neither requires nor entails having a self, whatever a self might be. As discussed below, the same holds for awareness and consciousness. In short, if an AI system manifests reflexivity that neither entails nor presupposes that it has a self.

[1] To illustrate, in their excellent paper, Lake et al. (2017) go over various key ingredients of human intelligence, and while they mention learning-to-learn and human-like cognitive flexibility which can be understood as requiring reflexivity, they do not mention reflexivity or any alternative ways to denote reflexivity, such as self-reference or self-representation. While there has been no explicit attention at artificially replicating the capacity for reflexivity, see (Johnson et al., 2024) for a recent call from the same direction as this perspective.

1. Reflexivity

What is reflexivity? The capacity for reflexivity allows an organism or system to refer to itself while registering that it is the object of reference. More technically:

> ***Reflexivity***: A system *S* has the capacity for reflexivity only if it can refer to (a part of) *S* while registering that it is *S*.

This definition of reflexivity mirrors the standard definition of first-person thought in linguistics and philosophy of language (Recanati, 2007).[2] In a representational system, reference and registration may amount to representation:

> ***Reflexivity***$_{representation}$: A system *S* has the capacity for reflexivity only if it can represent (a part of) *S* while representing that it is *S*.

For ease of presentation, we will assume that the relevant AI systems represent (a part of) *S*. However, in each case, the point can easily be reformulated while eschewing representationalist commitments. The qualification "while representing that it is *S*" rules out cases in which a system represents what happens to be (a part of) itself without registering that it is representing itself. A human example will help explain the difference between a system *S* representing *S* with and without registering that it is *S*. Say you are walking down a street and see the reflection of someone wearing a funny hat. As it happens, you are the person wearing the funny hat, but you are unaware that you are looking at yourself. Once you realize that you are the person in the mirror image, you realize that you are wearing a funny hat. First, you perceptually represent what happens to be you without exercising your capacity for reflexivity. In registering that you are the person wearing a funny hat, you represent yourself while manifesting reflexivity. In light of this specification, the manifestation condition on reflexivity can be specified as follows:

> ***Manifestation Condition on Reflexivity***: A system *S* manifests reflexivity only if it represents (a part of) *S* while representing that it is *S*.

The distinction between the capacity and its manifestation parallels the competence-performance distinction (Firestone, 2020). Indeed, a system or organism that performs an action or generates an output by luck or rote does not qualify as possessing the relevant capacity. To illustrate, consider someone who, for the very first time, uses a bow and arrow and hits the bull's eye but does so only due to a gust of wind, without which the arrow would have missed the target (Sosa, 2007). Such a person hits the bull's eye due to luck rather than having the relevant capacity. This raises the question of what it takes to qualify as possessing the capacity for reflexivity. There are at least two conditions on possessing a capacity: flexibility and aptness. To qualify as possessing a capacity, a system must manifest flexibility in employing it:

> ***Flexibility Condition***: A system *S* possesses capacity *C* only if it can successfully employ *C* across a range of relevant situations.

To illustrate, consider a system that has been trained to discriminate red from other colors. The system qualifies as possessing the capacity to discriminate red from other colors if it successfully discriminates not just the shades of red on which it was trained from other colors it encountered in training, but a range of red shades not encountered in training from a range of other colors. The flexibility condition is closely related to generalization (Lake & Baroni, 2023) and compositionality (Russin et al., 2024). The second condition is that the system can employ the capacity in relevant situations:

> ***Aptness Condition***: A system *S* possesses capacity *C* only if it can employ *C* successfully in relevant situations.

---

[2] It has been argued that at least some first-person thoughts do not refer to the individual entertaining the thought (Lewis, 1979). While such ways of understanding first-person content include the subject who produced the mental state in the content's index of evaluation, however, they do not manifest reflexivity. So, we can safely bracket such centered world views of *de se* content.

To illustrate, consider again the system trained to discriminate red from other colors. The system qualifies as possessing the capacity only if it employs that capacity to discriminate red when perceptually related to a red surface, where being related to a red surface means that (a) the system is spatially and temporally related to the red surface such that it can gain information about that surface via its sensory receptors and (b) the lighting conditions are such that it can detect the color of the surface. Applied to our topic here, the point is that a system only qualifies as possessing the capacity for reflexivity if it employs it flexibly and aptly.

How can this technical definition of reflexivity be operationalized to produce a system with the capacity for reflexivity? There are multiple ways in which an AI system could manifest reflexivity. Here are a few:

1. System *S* has a map or model of the world that explicitly represents (part of) *S*, allowing it to reidentify itself across space or time.
2. System *S* has the capacity for metacognition, allowing it to explicitly represent and manipulate its outputs while registering that *S* generated those outputs.
3. System *S* represents that two of its subsystems are both part of *S*.
4. System *S* represents that the outputs of two of its subsystems are both generated by *S*.

With all these options, reflexivity is in place. After all, with all these options, the system represents (part of) the system while registering that it is that system.

To get a better grip on the nature of reflexivity, it will be helpful to mention a few ways in which reflexivity manifests in humans before turning our attention to AI. They include self-awareness, introspection, reflexive source-monitoring, and metacognition. In self-awareness, we are conscious of ourselves (or some aspect of ourselves) from the first-person perspective. A particularly low-level form of self-awareness is proprioception, that is, the awareness of the position of one's limbs. In being inward-directed, self-awareness is distinct from being perceptually conscious of an object in one's environment. In other words, self-awareness is not a form of perception. Of course, a subject can look at her arm, but in doing so, she is aware of her arm via perception, not via inward-directed self-awareness.

It is important to note that an organism or system can manifest reflexivity without having the capacity for self-awareness or consciousness. So, self-awareness is sufficient but not necessary for reflexivity. This is crucial since—at the current state of development—AI systems arguably do not have consciousness. Nonetheless, they can have reflexivity. Thus, while exercising our capacity for reflexivity may manifest in self-consciousness, a system or organism can manifest reflexivity without having the capacity for consciousness. The converse holds equally. If pain qualifies as a form of consciousness, many animals have consciousness despite lacking reflexivity. In short, reflexivity and consciousness are doubly dissociated.

Introspection is a particularly high-level form of reflexivity. When introspecting, we are consciously aware of some aspect of ourselves (Morales, 2024). We can introspect on a preference, an emotional state, a belief, or a goal (Wu, 2023b). Source monitoring is the process by which an individual determines the provenance of a piece of information (Teng, 2024). The source could be vision, olfaction, memory, proprioception, testimony, or a book, to give just a few examples. If the source is internal to oneself, then monitoring that source will involve reflexivity. To be clear, perception, proprioception, and memory need not involve reflexivity. However, if one is tracking from which of those information sources a specific mental state stems, then one is exercising one's capacity for reflexivity. Finally, in metacognition, the capacity for reflexivity is applied to one's cognitive states (S. M. Fleming, 2024). This can take the form of thinking about one's reasoning, a belief one holds, a preference, or a goal. More generally, as discussed in more detail below, metacognition is a representation that is about a different representation of the same individual.

Introspection, source monitoring, and metacognition are all cognitively high-level forms of reflexivity. By contrast, self-awareness need not be high-level—though, naturally, only organisms with the capacity for consciousness can have this low-level form of reflexivity. Before turning our attention to reflexivity in AI systems, it will be helpful to note that reflexivity encompasses reflectivity, but not vice versa. Reflectivity involves taking a higher-order stance toward some aspect of oneself. By contrast, reflexivity does not require treating the system as the object of intentionality (Maiese, 2011). So, not all cases of reflexivity are cases of reflectivity.

2. Assessing if a System has Reflexivity: Mirror Tests and Mechanistic Interpretation

How can we assess whether a system manifests reflexivity? To address this question, let's first look at how reflexivity is tested in animals. The classic approach is the mirror test (Gallup et al., 2002): an unusual marking is painted on an animal's body without the animal noticing, and the animal is then placed in front of a mirror. It qualifies as passing the mirror test if it responds to its mirror image, for example, by shifting its gaze from its mirror image to its marked body part, by touching the marked body part, or by moving in front of the mirror in ways that allow it to get a better look at the unusual marking. By exhibiting such a response, the animal manifests that it registers that it is seeing itself in the mirror. Apes, dolphins, magpies, manta rays, elephants, ants, and numerous other animals, as well as humans at 18-24 months pass the mirror test (Gallup et al., 2002), though in some cases this is true only of a few specimens.

How could one replicate the mirror test for an AI system? Needless to say, the test need not include mirrors. The analogy to the mirror test would be that the system qualifies as having reflexivity if it responds to the detected anomaly in a way that implies that it registers that the anomaly is within its system. Consider a system that includes an anomaly detection, "self"-assessment, or "self"-diagnostic module. Such modules detect anomalies within the system, for example, by comparing current to past patterns. Suppose a system has detected an anomaly that happens to be in its network. How can we test whether the system registers that the anomaly is within its network? In other words, how do we distinguish whether the system is akin to the person who sees someone in a mirror with an unusual marking without realizing that it is that person (anomaly detection without reflexivity) or whether the system is akin to someone with the same visual input but who realizes that the marking is on her body (anomaly detection with reflexivity)?

Anomaly detection alone does not manifest reflexivity. However, the system may qualify as manifesting reflexivity if the anomaly detection module is combined with an adaptation module that appropriately responds to the anomaly detected such that the system manifests registering that the anomaly is within itself. That would replicate the difference between an animal seeing its mirror image without realizing that it is the animal with the marking (anomaly detection) and the animal realizing that it is looking at itself and manifesting this with an appropriate response directed at itself (anomaly detection plus appropriate response directed at the system). Machine learning programs that include a module that combines anomaly detection with adaptation are designed to assess their performance continually and adjust based on that assessment (Cretu-Ciocarlie et al., 2009). Such an adjustment can manifest reflexivity insofar as the algorithm operates, for example, on past outputs that it produced while representing that it made those outputs.

Now, the mirror test is famously imperfect for determining whether an animal has reflexivity. To illustrate, the animal could fail to pass the mirror test despite having the capacity for reflexivity because it does not care that it has a marking on its body (Plotnik et al., 2006). Moreover, pigeons pass the mirror test, but only after training (Gallup et al., 2002). This raises the question of whether pigeons have the capacity for reflexivity or if they can be trained to behave in ways that appear as if they do. In short, an animal may fail to pass the mirror test despite having the capacity for reflexivity (competence without performance) and may be trained to pass the mirror test despite lacking reflexivity (performance without competence). We can expect the same shortcomings by observing the behavior of AI systems.

Fortunately, we need not restrict ourselves to observable behavior to assess whether AI systems have reflexivity. An alternative is mechanistic interpretation, that is, the approach to understanding the mechanisms, internal architecture, and decision-making processes of machine learning systems, particularly deep neural networks (Kästner & Crook, 2024). Since reflexivity need not manifest in observable behavior, mechanistic interpretation will—in almost all cases—be necessary to determine whether a system has reflexivity. Mechanistic interpretation allows us to gain insight into the functional role of individual components of a neural network as well as the causal relationships between the components (Fleisher, 2022). Further, it allows us to gain insight into the flow of information through the network layers, specifically, how each component transforms the data (Conmy et al., 2023).

Among the many techniques used in mechanistic interpretation, ablation and activation methods would help assess whether a system has reflexivity. In ablation methods, a network component is removed or modified

to determine the functional role of the component in the network (Rai et al., 2024). In activation methods, the activation of components in response to different inputs is analyzed to determine their sensitivity to specific inputs (Bereska & Gavves, 2024). Jointly, ablation studies and activation analyses make it possible to break down a system's decision-making- and information-processing mechanisms into their components and determine each component's function as well as its relation to other components.

Before discussing existing cases of reflexivity in AI systems, it is important to note that several key features of AI systems do not manifest reflexivity even though their name might suggest otherwise. Most algorithms refer to parts of themselves via recursion or iteration. Neither recursivity nor iteration entails reflexivity. Similarly, a system with feedback loops need not manifest reflexivity. A feedback loop is a module in a system that operates on (part of) the system's output. So, the system's output is channeled back into the system as input for subsequent operations. The sheer existence of a loop is not sufficient for reflexivity. After all, reflexivity requires that the system register that it is the object represented. Iteration and recursivity, including feedback loops, need not include the system *S* representing (a part of) *S* while representing that it is *S*. In discussing existing cases of reflexivity in AI systems, we will contrast examples of systems that have reflexivity with closely related systems that do not. In doing so, we can determine what it takes for a system to manifest reflexivity rather than failing to do so.

3. Metacognition

We engage in metacognition when we think about our beliefs, evaluate our reasons, or reflect on our plans. More generally, metacognition can be specified as follows:

> ***Metacognition:*** A system *S* engages in metacognition only if *S* produces a metarepresentation R_r about a target representation *r* generated by *S* while registering that *S* generated *r*.

Note that this is a demanding understanding of metacognition. If we can show that AI systems have metacognition on this demanding notion, then it is easy to show that they have metacognition on less demanding notions. Since metacognition is always about one's own cognitive states, it is necessarily reflexive. The difference between metacognition and other manifestations of reflexivity is that the former is necessarily a second-order representation, while the latter could be a first-order representation. In metacognition, the system generates a representation of a distinct representation of the same system. In other manifestations of reflexivity, the system generates a representation of (part of) the system itself—rather than a representation this system produced.

To assess whether AI systems manifest metacognition, we will focus on Large Language Models (LLMs). However, AI systems other than LLMs may have metacognitive capacities. Moreover, as argued below, LLMs manifest reflexivity in ways that do not take the form of metacognition. Regardless of the notion of cognition in play (Barack & Krakauer, 2021), it is, of course, controversial whether LLMs have cognitive capacities, let alone metacognitive ones (Shiffrin & Mitchell, 2023). For the sake of argument, let's assume that the capacities they employ to generate sentential output qualify as cognitive capacities (for support, see (Pavlick, 2023)).

There are multiple levels at which to address the question of whether LLMs have metacognition. First, LLMs can produce sentences that seem to express that they engage in metacognition, such as sentences of the form "I believe that *p.*" The fact that they can generate such sentences does not entail that they have the metacognitive capacities such sentences seemingly express. After all, generating such sentences can be due to a learned pattern of language use rather than a manifestation of a capacity.

An analogy will help explain why. An AI therapist can produce sentences that seemingly express empathy. However, this does not entail that it has empathy or any other emotional states. An AI therapist merely has the cognitive capacity to produce sentences that seemingly express emotions. It does not have the emotions those sentences seemingly express. Indeed, it lacks the hardware required to have emotions. Consequently, it cannot feel empathy.

An LLM that produces sentences seemingly manifesting metacognitive capacities can be understood analogously. As LLMs do not have the emotions seemingly expressed with the sentences they output, they may

not have the reflexive capacity they seem to manifest when they output sentences of the form "I believe that *p*." While such sentences have the form of metacognitive thoughts, they may not express such thoughts. So, the fact that LLMs can generate sentences that seem to manifest metacognition does not cut any ice as to whether LLMs have the capacity for reflexivity.

The critical dimension on which to assess whether an LLM has reflexivity is whether it represents or operates on content it produced while registering that it produced that content. That would qualify as taking a metacognitive stance to the content it produced rather than merely stochastically parroting sentences that superficially have the form of metacognition. "Self"-attention mechanisms in LLMs fit the bill. Importantly, however, the notion of attention in play has nothing to do with consciousness.

Self-attention mechanisms allow an LLM to weigh the relevance of different parts of a conversation as it generates new contributions to that conversation (Buckner, 2023). While responding to the user's input, the self-attention weights represent, at each step, what the LLM considers most relevant from its response so far. Specifically, each input sentence is prefixed with a special self-attention token that serves as a bottleneck: it forces the model to selectively compress and prioritize the relevant aspects of its prior output (Luo et al., 2023). In doing so, the model generates representations that operate on these outputs. Moreover, the mechanism allows the LLM to summarize and comment on its prior outputs in the conversation. If we assume that LLMs have cognitive capacities, then these representations are a form of higher-order cognition. After all, they operate on other representations produced by the same system.

Now, for this to qualify as metacognition, it is not sufficient for the LLM to generate representations that summarize or comment on its previous outputs; the LLM must register that it generated those outputs. The attention mechanism represents which part of the conversation was generated by itself. So, the LLM registers the source of each previous sentence. With this addition, the higher-order representations have all the hallmarks of metacognition: the LLM not only generates sentential output while operating on its previous output; it registers that it produced that output. This process of prioritizing some of its outputs over others provides the LLM with a dedicated computational pathway for representing and commenting on its previous outputs. To illustrate, suppose an LLM is asked about the French Revolution, and it starts with producing sentences about enlightenment and freedom. Halfway through the response, the self-attention mechanism registers that its response so far primarily consisted of sentences about enlightenment and freedom. In response, the LMM may shift to other topics, such as the social, political, and economic factors that led to the French Revolution, the ancien régime, or the role of the Jacobins.

In this way, the metacognitive representation of its past output informs the LLM's text-generation. Crucially, these higher-order representations are enabled by the LLM's self-attention mechanism, which allows it to integrate information from its previous outputs. This supports the thesis that the LLM is engaging in metacognition and not simply replicating a superficial language pattern. Indeed, the mechanism allows the LLM to exhibit a form of "self"-control over its text-generation process. Such self-control is a hallmark not just of metacognition but of reflexivity more generally.

A specific form of metacognition is metareasoning, that is, reasoning about one's reasoning. An LLM that can manipulate its inference strategies, decision-making procedures, or problem-solving approaches does not necessarily manifest metareasoning capacities. Only if the manipulation amounts to reasoning would the LLM qualify as engaging in metareasoning. An LLM that includes a rationale generation module (Ehsan et al., 2018) could be argued to have the capacity for metareasoning. The architecture of such an LLM includes an additional "rational generation"-step. The model first generates a reason for a claim and then provides a reason for the initial reason, for example, by evaluating its quality and coherence. This is achieved by training the LLM not only to generate textual outputs but, moreover, to generate a reason for why that output is appropriate or optimal.

Other candidates for metacognition in LLMs include self-interpretation (Chen et al., 2024), self-knowledge (Kadavath et al., 2022), and introspection (Binder et al., 2024), among others. It would lead too far afield to discuss each of these in detail here. Importantly, while the focus here was on metacognition in LLMs, AI systems other than LLMs may have metacognitive capacities, and not all manifestations of reflexivity in LLMs amount to metacognition. As discussed below, chain-of-thought reasoning can lead a model to exhibit second-

guessing behavior. Chain-of-thought reasoning prompts an LLM to break down problems into intermediate steps before generating a response to a question (Wei et al., 2023). Second-guessing is a manifestation of reflexivity in LLMs that need not amount to metacognition or metareasoning.

To conclude the discussion, we note that while LLMs manifest metacognition, they do not have a more generalized capacity for reflexivity. Such a capacity would be useful, specifically for LLMs. LLMs are excellent at detecting inconsistent content. Not only can they detect such content, they can explain that inconsistencies violate fundamental laws of logical reasoning and are to be avoided. Nonetheless, they generate such content, at least in their current stage of development. Adding a reflexive module would be a way to fix this problem. Problems of this kind can be fixed by building in guardrails. To illustrate, most LLMs no longer produce blatant racist or sexist content. The reason is that they have built-in guardrails stopping them from generating such content. Adding a module with reflexive capacities to an LLM would allow it to refrain from generating any content it deems problematic without the piecemeal approach of adding guardrails after problems arise. A general capacity for reflexivity would allow an LLM to exhibit "self"-control with regard to the content it generates in ways that mimic how we monitor ourselves.

4. Meta-reinforcement learning

A reinforcement learning module provides feedback to a system in the form of rewards or penalties, where information about rewards (or penalties) is provided to the network in terms of meeting (or failing to meet) thresholds. By adjusting its strategy to maximize rewards, the system uses this feedback to achieve a goal. A simple reinforcement learning module need not have the capacity for reflexivity if it is merely a feedback loop with a twist, namely a reward (or a penalty). There is nothing reflexive about meeting a threshold. So, the twist to a feedback loop does not amount to a manifestation of reflexivity.

However, a meta-reinforcement learning module (Meta-RL) can be understood as manifesting reflexivity. Meta-RL allows a Recurrent Neural Network (RNN) to learn how to optimize its learning algorithms. This enables the system to adapt to new tasks by updating its learning process based on its interactions with the environment (Wang et al., 2017). This is achieved (i) by training the RNN on a distribution of tasks and optimizing the RNN's parameters to maximize the expected cumulative reward over all tasks. A second key component of Meta-RL is (ii) a meta-learner network. This network learns an update rule that it applies to update the RNN's parameters. More specifically, the meta-learner takes the system's current parameters and their loss gradients as input, based on which it then updates the system's parameters (Finn et al., 2017). By learning to generate parameter updates, the meta-learner can adapt the system's learning algorithm to new tasks with only minimal exposure to relevant additional data (Ravi & Larochelle, 2017).

To illustrate with an example, consider RL^2 (Duan et al., 2016). The RNN takes its current state, previous output, and previous reward as input, on the basis of which it generates a representation that encodes its past interactions with the environment. In performing further tasks, the system then updates this reflexive representation and uses this updated representation to make decisions about future tasks. This representation has all the hallmarks of a reflexive representation. After all, it encodes information that is not only a function of a current representation based on the system's inputs, outputs, and their consequences; it depends also on its learning process. The learned learning algorithm is encoded in the RNN and adapts to new environments by operating on its reflexive representations to learn new task-specific functions.

More generally, a meta-RL module refers to the target module by modifying its architecture. Thereby, the system adapts its structure and learning algorithms based on its interactions with the environment. Since the meta-RL and the target module are part of the same system, the architecture modifications are a manifestation of reflexivity.

5. Self-Improving Systems

So-called "self"-improving AI systems are designed to improve themselves over time autonomously, much like the human brain continuously adapts. In a trivial way, any AI system that refines its feature map in response to

new data, thereby ameliorating its predictions, is self-improving. However, improving a feature map in the face of new data, thereby ameliorating predictions, does not on its own exhibit anything that qualifies as reflexivity.

Examples of self-improving systems that manifest reflexivity include Self-Modifying Neural Networks and Attention-Based Recurrent Neural Networks. Meta-RL can be understood as a special kind of self-improving system. However, since the neural network learns to optimize its learning process rather than directly improving its performance on a specific task, it is best categorized as a different type of reflexivity. Let's take a closer look at Self-Modifying and Attention-based RNNs. A Self-Modifying Neural Network modifies its architecture to improve its performance on a given task (Schmidgall, 2020). The network's performance on the task depends on its architecture, and the reinforcement learning algorithm modifies this architecture based on observed performance (Real et al., 2020). In other words, the system makes structural changes to itself as a response to the consequences of its past architectural decisions that it registers as its own. Adaptations to its architecture can amount to adding or removing network layers, adjusting the number of weights, changing its hyperparameters, or modifying the connections between weights. The reinforcement learning algorithm acts as a "controller" that learns to modify the network's architecture to maximize its performance on the task. The critical point for the current discussion is that the network's current architecture is taken as input to generate modifications to this architecture. Consequently, the neural network manifests reflexivity in that it encodes its architecture and updates and optimizes this architecture by interacting with a reinforcement learning algorithm.

In an Attention-Based RNN, the network learns to attend to its previous hidden states and outputs when processing new inputs and generating new outputs (Kim et al., 2025). The attention in play radically differs from human attention in that it does not include awareness or consciousness (Wu, 2023a). Due to its attention mechanism, the system can selectively refer to relevant parts of its computation history. More specifically, the Attention-Based RNN computes an attention distribution over its relevant previously hidden states, which it uses to weigh and combine those states into a context vector (Vaswani et al., 2023). This context vector is then used as an input to compute the new hidden state and output. Attention-based RNNs can be understood as manifesting reflexivity since the RNN operates on its internal representations (hidden states) while registering, via the attention weights, that those representations are its own. Thereby, the network uses its own processed information to inform its subsequent processing.

6. Second-Guessing: The Dark Side of Reflexivity

Reflexivity has a dark side. In humans, it can lead to paralyzing self-doubt. Recently, similar phenomena have been observed in LLMs (Khan et al., 2023), specifically in the context of chain-of-thought reasoning (CoT). CoT prompts an LLM to decompose a question or problem into smaller units (Wei et al., 2023). This has many benefits that mimic the benefits of second-guessing. It allows an LLM to consider multiple possible solutions to a problem before settling on one approach. Second, decomposition enables an LLM to detect errors midway through generating text, retrace its steps, and correct the errors (He et al., 2025). Third, by decomposing a problem, an LLM can reconsider assumptions made and restart answering a question with alternative assumptions. Fourth, it allows an LLM to register when an approach is leading to undesirable outcomes and pivot to a different approach (Creswell et al., 2022). An LLM might start solving a problem in one way, find that it leads to difficulties, and respond by reorienting and trying an alternative strategy.

These capacities can lead an LLM to exhibit behavior similar to a self-critical researcher who carefully thinks through her approach to a problem and questions her assumptions. Such self-doubt may be the ultimate sign of human-level intelligence (S. Fleming, 2020). However, it is not ideal if self-doubt traps a researcher in an endless loop of questioning assumptions and pivoting between multiple possible strategies.

As in humans, second-guessing in LLMs can lead to improved outcomes but also to trouble. In evaluating different possible approaches, LLMs can get into loops of excessive verification (Khan et al., 2023). Such overthinking significantly increases computational load without proportional benefits in outcome (Zhou et al., 2025). More problematically, second-guessing can lead a model to retract initial correct answers and flip to incorrect answers after lengthy and resource-costly verification of the initial correct answer (Liu et al., 2024). This can be a result of the LLM focusing on irrelevant aspects of the problem after decomposition.

The critical question is how to prioritize the conflicting goals of developing AI systems that have human-level intelligence—including second-guessing and self-doubt, on the one hand, and AI systems that are efficient, on the other. Short of self-doubting paralysis, an Al system that questions its approach and assumptions at the cost of being slower is arguably preferred. Khan et al. (Khan et al., 2023) suggest that if LLMs selectively use second-guessing, its pitfalls can be avoided while reaping its benefits. Applying decomposition indiscriminately to all questions can lead to excessive verification and focus on irrelevant aspects of a problem. However, as Khan et al. (Khan et al., 2023) argue, this problem can be avoided by restricting question decomposition to cases in which the LLM has low confidence in an answer: The model first attempts to answer the question without decomposing it. Then, it checks its confidence in its answer against a provided threshold. If the confidence score does not meet the threshold, the decomposition process is triggered. Second-guessing selectively when confidence is low mimics the way humans allocate their attention and cognitive resources to problems where certainty is low, and the risks are high.

7. Conclusion

We analyzed four ways current AI systems manifest reflexivity: metacognition, metalearning, self-improvement, and second-guessing. In each case discussed, the relevant systems satisfy the flexibility and aptness conditions. After all, in each case, the relevant system exhibits considerable range and exercises the capacity successfully in situations in which the employment of the capacity is warranted. It is important to note, however, that each of these manifestations of reflexivity is independent of the other.

An objection waiting in the wings is whether the same capacity for reflexivity is manifested in each of the four cases. In response, no doubt, metacognition, metalearning, self-improvement, and second-guessing are each distinct manifestations of reflexivity. However, reflexivity is the common denominator. To explain, recall that a system *S* has the capacity for reflexivity only if it can represent (a part of) *S* while registering that it is *S*. Recall also that it *manifests* that capacity only if it represents (a part of) *S* while registering that it is *S*. As argued, in each case discussed, the relevant system represents (a part of) *S* while registering that it is *S*. So, while the manifestation of reflexivity is distinct in each case, the condition for manifesting reflexivity is satisfied. In short, reflexivity is the common denominator of metacognition, metareasoning, self-attention, self-improving systems, metalearning, and second-guessing, to mention just a few forms of reflexivity discussed in this perspective.

While the relevant systems satisfy the flexibility and aptness condition for the specific type of reflexivity in each case discussed, the gap between humans and AI systems is vast with regard to the range and flexibility in employing the capacity. This is not surprising. After all, no explicit attention has been directed at artificially replicating this key element of human intelligence. With more research into artificially replicating the capacity for reflexivity, this gap could be closed.

References

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), 359–371.

Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety—A Review. https://doi.org/10.48550/arXiv.2404.14082

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., & Evans, O. (2024). Looking Inward: Language Models Can Learn About Themselves by Introspection. https://doi.org/10.48550/arXiv.2410.13787

Buckner, C. J. (2023). *From Deep Learning to Rational Machines*. Oxford University Press.

Chen, H., Vondrick, C., & Mao, C. (2024). SelfIE: Self-Interpretation of Large Language Model Embeddings. https://doi.org/10.48550/arXiv.2403.10949

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. https://doi.org/10.48550/ArXiv.2304.14997

Creswell, A., Shanahan, M., & Higgins, I. (2022). Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. https://doi.org/10.48550/arXiv.2205.09712

Cretu-Ciocarlie, G. F., Stavrou, A., Locasto, M. E., & Stolfo, S. J. (2009). Adaptive Anomaly Detection via Self-calibration and Dynamic Updating. In E. Kirda, S. Jha, & D. Balzarotti (Eds.), *Recent Advances in Intrusion Detection,* pp. 41–60.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL2: Fast Reinforcement Learning via Slow Reinforcement Learning.

Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2018). Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 81–87.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *Proceedings of the 34th International Conference on Machine Learning*, 1126–1135.

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

Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme*, *19*(4), 534–560.

Fleming, S. (2020). *Know Thyself*. Hachette.

Fleming, S. M. (2024). Metacognition and Confidence: A Review and Synthesis. *Annual Review of Psychology*, *75*(Volume 75, 2024), 241–268.

Gallup, G., Anderson, J., & Shillito, D. (2002). The mirror test. In *The cognitive animal: Empirical and theoretical perspectives on animal cognition* (pp. 325–333). MIT Press.

García-Carpintero, M. (2024). The *Real Guarantee* in *De Se* thought: How to characterize it? *The Philosophical Quarterly*, pqae133.

He, Y., Li, S., Liu, J., Wang, W., Bu, X., Zhang, G., Peng, Z., Zhang, Z., Zheng, Z., Su, W., & Zheng, B. (2025). Can Large Language Models Detect Errors in Long Chain-of-Thought Reasoning? https://doi.org/10.48550/arXiv.2502.19361

Johnson, S. G. B., Karimi, A.-H., Bengio, Y., Chater, N., Gerstenberg, T., Larson, K., Levine, S., Mitchell, M., Rahwan, I., Schölkopf, B., & Grossmann, I. (2024). Imagining and building wise machines: The centrality of AI metacognition. https://doi.org/10.48550/arXiv.2411.02478

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., … Kaplan, J. (2022). Language Models (Mostly) Know What They Know. https://doi.org/10.48550/arXiv.2207.05221

Kästner, L., & Crook, B. (2024). Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science*, *14*(4), 52.

Khan, Z., BG, V. K., Schulter, S., Chandraker, M., & Fu, Y. (2023). Exploring Question Decomposition for Zero-Shot VQA. https://doi.org/10.48550/arXiv.2310.17050

Kim, D., Tanwar, S., & Kang, U. (2025). Accurate multi-behavior sequence-aware recommendation via graph convolution networks. *PloS One*, *20*(1), e0314282.

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115–121.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Lewis, D. (1979). Attitudes De Dicto and De Se. *Philosophical Review*, *88*(4), 513–543.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024). Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse. https://doi.org/10.48550/arXiv.2410.21333

Luo, Q., Zeng, W., Chen, M., Peng, G., Yuan, X., & Yin, Q. (2023). Self-Attention and Transformers: Driving the Evolution of Large Language Models. *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, 401–405.

Maiese, M. (2011). *Embodiment, Emotion, and Cognition*. Palgrave.

Morales, J. (2024). Introspection Is Signal Detection. *The British Journal for the Philosophy of Science*, *75*(1), 99–126.

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), 20220041.

Plotnik, J., Waal, F., & Reiss, D. (2006). Self-Recognition in an Asian Elephant. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 17053–17057.

Rai, D., Zhou, Y., Feng, S., Saparov, A., & Yao, Z. (2024). A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models. https://doi.org/10.48550/arXiv.2407.02646

Ravi, S., & Larochelle, H. (2017). Optimization as a Model for Few-Shot Learning. *Proceedings of the International Conference on Learning Representations*.

Real, E., Liang, C., So, D., & Le, Q. (2020). AutoML-Zero: Evolving Machine Learning Algorithms From Scratch. *Proceedings of the 37th International Conference on Machine Learning*, 8007–8019.

Recanati, F. (2007). *Perspectival Thought: A Plea for (Moderate) Relativism*. Oxford University Press.

Russin, J., McGrath, S. W., Williams, D. J., & Elber-Dorozko, L. (2024). From Frege to chatGPT: Compositionality in language, cognition, and deep neural networks. https://doi.org/10.48550/arXiv.2405.15164

Russin, J., O'Reilly, R. C., & Bengio, Y. (2020). Deep Learning needs a prefrontal cortex. *International Conference on Learning Representations (ICLR)*.

Schellenberg, S. (2025). Polysemy of "I." *Mind & Language*.

Schmidgall, S. (2020). Adaptive Reinforcement Learning through Evolving Self-Modifying Neural Networks. https://doi.org/10.48550/arXiv.2006.05832

Shanahan, M. (2024). Talking about Large Language Models. *Commun. ACM*, *67*(2), 68–79.

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, *120*(10), e2300963120.

Sosa, E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Oxford University Press.

Teng, L. (2024). The Epistemic Insignificance of Phenomenal Force. *Philosophy and Phenomenological Research*, *109*(1), 55–76.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2017). Learning to reinforcement learn. https://doi.org/10.48550/arXiv.1611.05763

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Wu, W. (2023a). *Movements of the Mind: A Theory of Attention, Intention and Action*. Oxford University Press.

Wu, W. (2023b). On Possible and Actual Human Introspection. *Journal of Consciousness Studies*, *30*(9–10), 223–234.

Zhou, X., Tie, G., Zhang, G., Wang, W., Zuo, Z., Wu, D., Chu, D., Zhou, P., Sun, L., & Gong, N. Z. (2025). Large Reasoning Models in Agent Scenarios: Exploring the Necessity of Reasoning Capabilities. https://doi.org/10.48550/arXiv.2503.11074