Reflective Artificial Intelligence

Peter R. Lewis[1*†] and Ștefan Sarkadi[2*†]

[1*]Ontario Tech University, Oshawa, Ontario, Canada.
[2*]Department of Informatics, King's College London, UK.

*Corresponding author(s). E-mail(s):
peter.lewis@ontariotechu.ca; stefan.sarkadi@kcl.ac.uk;
†These authors contributed equally to this work.

Abstract

As Artificial Intelligence (AI) technology advances, we increasingly delegate mental tasks to machines. However, today's AI systems usually do these tasks with an unusual imbalance of insight and understanding: new, deeper insights are present, yet many important qualities that a human mind would have previously brought to the activity are utterly absent. Therefore, it is crucial to ask which features of minds have we replicated, which are missing, and if that matters. One core feature that humans bring to tasks, when dealing with the ambiguity, emergent knowledge, and social context presented by the world, is reflection. Yet this capability is completely missing from current mainstream AI. In this paper we ask what *reflective AI* might look like. Then, drawing on notions of reflection in complex systems, cognitive science, and agents, we sketch an architecture for reflective AI agents, and highlight ways forward.

Keywords: reflective AI, reflection, meta-reasoning, agent architectures

1 Introduction

Margaret Boden has described artificial intelligence as being about making 'computers that do the sorts of things that minds can do' (Boden, 2016, p1). One strength of this definition lies in the fact that it does not start from an arbitrary description of the things that might be necessary or sufficient for a system to 'count' as AI, but it encourages us to ask: what are the sorts of things

that our minds do? Further, a curious mind is then tempted to ask: could we replicate these things? If so, how? If not, why not and does that matter?

The definition also implies that there are things that human minds currently do, that in the future machines might do instead. This is not only true now, but as Mayor (2018) discusses, has been the case since antiquity and likely will be far into the future. It is this transference of activity that gives rise to the seemingly constant stream of examples of new 'AI technologies'. These mostly do things that human minds used to, or wished to do. This is, of course, also the source of many of the issues and benefits that arise from the creation and use of AI technology: as we figure out how to replicate some of the things that minds can do, we delegate these things to machines. This typically brings increased automation, scale, and efficiency, which themselves contain the seeds of both enormous potential social and economic benefit, and potential real danger and strife.

Further, we can notice that these AI technologies usually do this with an unusual (im)balance of insight and understanding. New, deeper insight and understanding often arise from the models employed, while many of the 'qualities' that a human mind would have previously brought to the activity, are utterly absent.

In designing and analysing embodied AI technologies, the concept of an *intelligent agent* is central, and necessitates descriptions that are abstracted from the natural intelligences they are inspired by or modelled on. This abstraction, in turn, means that the notion of an AI agent only partially captures the mental, cognitive, and physical features of natural intelligence. Hence, it is important to ask: are the features that we have included sufficient for what is needed? Are we satisfied with leaving out those which we did?

Frank and Virginia Dignum have recently reminded us how powerful the concept of an agent is in AI (Dignum and Dignum, 2020). They have also pointed out some of the paradigmatic failures of *agent-based modelling* (ABM) and *multi-agent systems* (MAS). ABM methodologies aim to describe a large and complex system of agent populations by using analysis tools in the form of agent-based simulation. MAS methodologies, instead, focus on the operational side of interacting systems, where agents operate to create changes in their environment. However, neither methodology is fit for designing human-like agent architectures. The focus of their discussion (Dignum and Dignum, 2020) is to propose a social MAS architecture and argue that future socially-aware AI architectures should be different from today's common utility- and goal-driven models. A similar proposal was made by Ron Sun years ago, namely that agent methodologies need Cognitive Science and vice-versa (Sun, 2001) to address complex socially-aware AI and be able to design such architectures. Antonio Lieto refreshes this proposal (Lieto, 2021). In this paper we continue this line of thought, sketching an agent architecture that captures some reflective capabilities, based on cognitive theories. Due to the complex and modular nature of reflection it is impossible to find a single unique and crisp clear definition of the term reflection (Pitt, 2014). Reducing the definition to a single

process or component of an architecture would fail to address the richness of this cognitive process and would be counter-productive in explaining how all of the processes and components at play interact for human-like reflection to happen. Thus, in order to do it justice, similarly to (Tine, 2009), we adopt a differentiated theory approach to convey the notion of reflection.

2 Playing Chess Isn't Just About Chess

So what are the sorts of things minds do? As Richard Bellman suggested in 1978 (Bellman, 1978), these include activities such as 'decision-making, problem solving, learning, creating, game playing, and so on.'[1] And the sheer quantity of research on machines that can do these activities is astounding. Yet as Bellman's 'and so on' suggests, this is clearly an incomplete list. Perhaps any such list would be. It might be more useful to think situationally. We can ask: which features of our minds do we bring to different activities? Let us explore a thought experiment using a canonical example: chess. When playing chess, we largely bring the ability to reason, to plan ahead, to use heuristics, and to remember and recall sequences of moves, such as the caro-kann defence. Against an anonymous opponent on the Internet, we might try to use these abilities as best we can.

When playing chess with a child, however, we might typically bring a few more features too: patience, empathy (for example to understand the child's current mental model of the game to help coach them), and also some compassion, since proficient chess players could likely beat most children every time and make it less interesting all round. Letting children win is also not helpful, but a parent might play out a different sequence of moves to open up more in-game experiences from time to time. As a young player grows up, benefiting from both more brain development and experience at chess, and finds joy in different parts of the game, the way an adult opponent might do this will change. A good teacher might think back over previous games, reflect on the changes in the child's understanding and reasoning, and responses to moves. They might use this to speculate on and mentally play out possible future games.

This chess example illustrates three points: (i) even playing chess is not just about problem solving; (ii) rather unsurprisingly, our mental features are rich, contextual, and flexible; and (iii) we reflect on our situations, our current and past behaviour in them, and the likely outcome of those behaviours including the impact on others, in order to choose which mental features to engage. This is not just about flexible behaviour selection, it's about which mechanisms – which of Boden's (1998) 'sorts of things' – even kick in. What can this teach us about how we might want to build AI systems? Returning to the idea that we are delegating mental activity to machines, it tells us that perhaps we might want to have a similar ability in AI agents.

[1] Interestingly, this was shortened to simply 'decision-making, problem solving, learning...' by Russell and Norvig (2021), and it is this truncated version with the ellipsis that is most commonly quoted.

3 The Dangers of Incomplete Minds

Many people tie themselves in knots trying to define 'intelligence', hoping that that will lead us to somewhat of a more complete (and, they often say, more helpful) definition of 'artificial intelligence'. One example of such a discussion can be found in a recent special issue of the *Journal of Artificial General Intelligence* (Monett et al, 2020). As pointed out by Sloman in that collection, much of this definitional wrangling misses the point, at least from the perspective of deciding when we want to accept a computer to replace part of the activity previously done in society by human minds. Better questions might be to ask: what can this thing *do*; and what is it *for*? Consider: if we are deciding to put a machine in a position where it is carrying out a task in a way that we are satisfied is equivalent to what previously only a human mind could do, we have admitted something about the nature of either the task, or the machine, or our minds. Perhaps an AI system is simply a machine that operates sufficiently similarly to our mind, at least in some situations, that we are prepared to accept the machine operating in lieu of us. So this leads us to ask when and why we would be prepared to accept this. Or perhaps, given most AI systems (and minds) cannot be not fully understood or controlled, when and why we would be prepared to trust it to do so (Lewis and Marsh, 2021).

In one recent example, the seemingly harmless act of allowing a 'smart' voice assistant to propose entertainment activities to a child led to a life-threatening suggestion from a supposed trusted AI[2]. Normally, when delegating the proposal of children's play activities, we would expect that the person we had delegated that to would have not only a decent dose of common sense, but also the ability to consider the potential consequences of any ideas that sprung to mind before vocalizing them.

In another now well-known example, Amazon's automated recruiting tool, trained on data from previous hiring decisions, discriminated based on gender for technical jobs (Reuters, 2018). Here, the delegation is from professional recruiters and hiring managers to a computer that replicates (some of) the mental activity they used to do. The aims are automation, scale, and efficiency. That such a sexist system was put into practice at all is at the very least unfortunate and negligent. It is also tempting to argue that these are 'just bad apples', and that better regulation is the answer. It may be, but even then it is likely to be insufficient (Powers et al, 2023). But what is particularly interesting in our context is that people – hiring managers, shareholders, applicants – trusted the system to do something that, previously, a human mind did. But unlike the mind of the professional it replaced, it had no way of reflecting on the social or ethical consequences, or on the virtue or social value of its actions, or even if its actions were congruent with prevailing norms or values. That it had no way of reflecting on this meant that it also stood no chance at stopping or correcting itself. Indeed, neither of the above AI systems even had the mental machinery to do such a thing – this part of the mental activity is, as

[2] https://www.bbc.co.uk/news/technology-59810383

yet, nowhere near delegated. This leads to an unusual divorce of accompanying mental qualities that would normally work in concert. No wonder the behaviour might seem a little pathological.

As humans, a core part of our intelligence is our ability to reflect in these ways; reflection is a core mental mechanism that we use to evaluate ourselves. The existence of this form of self-awareness and self-regulation can be key to why others may find us trustworthy. Could we expect the same of machines?

4 The Role of Reflection in Driving Human Behaviour

One aspect of reflection is captured by what Socrates called his 'daemon' (Nesselrath et al, 2010), something that checked him 'from any act opposed to his true moral and intellectual interests' (Plato (translated by Paul Shorey), 1969). Socrates saw this as a divine signal, not proposing action, but monitoring it, and intervening if necessary. If such a check were based on morals or ethics, we might call this a conscience. If it were based on broader goals than simply the immediate (for example, choosing a chess move to make against your daughter), we might call this considering the bigger picture. Essentially, this is a process that notices what we are thinking, what we are considering doing, and allows and explores the thought, but can prevent the action. It decides whether to do this by contextualising the action. Contexts, as alluded to above, might be ethical, cultural, political, social, or based on non-immediate (higher-level, longer-term, or not immediately visible) goals.

What Socrates presents here requires a 'Popperian' mind according to Dennett's Tower of Generate and Test (Dennett, 2013, 2008, 1996). Essentially, in what he describes as a framework for 'design options for brains', Dennett notes that (at the bottom of the Tower) the testing of hypotheses is done by Darwinian evolution: hypotheses are generated through mutations and the placing of novel organisms in the world and tested through their survival. Above this, Skinnerian creatures test hypotheses by taking actions and learning in an operant conditioning fashion, based on environmental feedback within their lifetime. Higher still are Popperian and Gregorian creatures[3], which have the mental capability to bring hypothesis testing internally to their mind, rather than requiring it to be done in the world. Both of these operate with forms of reflection: put simply, Popperian creatures think about what to think, and Gregorian creatures, using tools, language, and culture, extend this to think about *how* to think.

One plausible way these Popperian and Gregorian creatures' minds might work is Hesslow's Simulation Theory of Cognition (Hesslow, 2002, 2012). Hesslow's hypothesis is that there exists a mechanism in the brain that helps agents reason about the consequences of their actions in an environment by simulating the stimuli of their behaviour in that environment, without having

[3]Dennett adopted the term *'Gregorian creatures'* based on the British psychologist Richard Gregory (Dennett, 1996, p99).

this behaviour previously reinforced by actual stimuli generated by their past behaviour. For example, this mechanism allows an agent to think about the deadly consequences of driving towards a concrete wall at high speed without having done it beforehand.

Schön (1984) compares the inherent nature of reflection in professional practice with a purely technically rational approach that might be characterised by up-front specification and subsequent problem solving. As opposed to passive problem solving, Active Experimentation is emphasised by how professional practitioners deal 'on-the-fly' with uncertainty, ambiguity and emergent knowledge inherent in tasks. From a technical rationality perspective, Schön (1984) argues, 'professional practice is a process of problem solving', yet, 'in real-world practice, problems do not present themselves to the practitioner as givens. They must be constructed from the materials of problematic situations which are puzzling, troubling, and uncertain.' This means that the sorts of things that (to continue the example) a professional recruiter does is not simply problem solving in a defined setting: there are patterns and mechanical aspects to their work, but the problem is always somewhat uncertain, and emerges from practice and the setting. Thus, we arrive at what Schön describes as 'an epistemology of practice which places technical problem solving within a broader context of reflective inquiry.'

Similarly, Weinberg (1972) argues that scientific knowledge and problem solving take place within a broader, untidier, and chaotic complex world. This world contains questions that, although appearing scientific, in fact transcend science in their nature. It is here where reflection based on experience can be a mechanism for contextualizing operational knowledge and problem solving within this trans-scientific world.

A model that captures reflection in practice, that is both exploratory and governed by a sense of the bigger picture and the principles that govern our intended direction, is Kolb's learning cycle (Kolb, 1984). His Experiential Learning Model (illustrated in Figure 1) comprises four phases: i) having a concrete experience, ii) an observation and subjective evaluation of that experience in context, iii) the formation of abstract concepts based upon the evaluation, and iv) the formation of an intention to test the new concepts, leading to further experience.

5 Where are we now?

Reflection in humans is complex and comprises numerous related phenomena. This makes it extremely difficult, if not impossible to find a single, crisp, and clear definition. This is usually the case with complex socio-cognitive phenomena (e.g., Pitt (2014)). Our approach instead is to contribute to building a 'differentiated theory' (Tine, 2009), as is often done in social psychology. This allows us to collect and compare the different ways in which phenomena all commonly referred to as being part of 'reflection' interact. In doing so, we

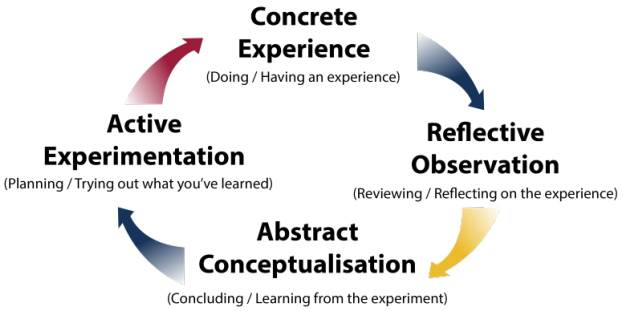

Fig. 1 Kolb's Experiential Learning Model. Source: (Kolb, 1984). The model shows captures the cognitive cycle in humans that is responsible for learning from experience.

aim to build towards a socio-cognitive theory of reflection in AI. Let us first examine the current state of AI, in this light.

Critic Agent Architecture

Introduced by Russell and Norvig (2021) and considered to be 'the' mainstream AI agent architecture due to its source's ubiquity in most AI courses taught around the world, the Critic Agent architecture represents the relation between the modules responsible with a learning agent's behaviour. According to Russell and Norvig (2021), the *performance element* is responsible for selecting actions the agent performs in the environment, and can itself be considered a basic agent architecture, e.g. a reflex agent or a model-based agent or a utility agent which is responsible for decision-making. What makes the Critic Agent different from other mainstream agent types is its ability to learn and operate in unknown environments due to the *critic*, *learning element* and *problem generator* modules. The *learning element* enables the agent to learn from feedback provided by the *critic* module after evaluating the rewards and penalties interpreted by the *performance standard*, while the *problem generator* is responsible for suggesting actions that lead to new information, e.g. responsible with guiding exploration of an unknown environment.

The Critic Agent is arguably the most 'advanced' and certainly the most complete architecture proposed by Russell and Norvig (2021), and hence we use it as a starting point in our analysis. Despite the explicit articulation of the processes of learning, obtaining feedback, and exploration of new solutions, there is nothing present in the architecture that captures the notion of reflection as discussed above. Considering the Critic Agent allows us to see that something additional is needed.

Artificial Neural Network Architectures (ANN)

Initially introduced by McCulloch and Pitts (1943) and in the form of a simple perceptron by Rosenblatt (1958), ANNs have gained major traction in the AI community. ANNs are highly applicable in the domain of statistical machine learning in which they are trained to perform various tasks, and outperform

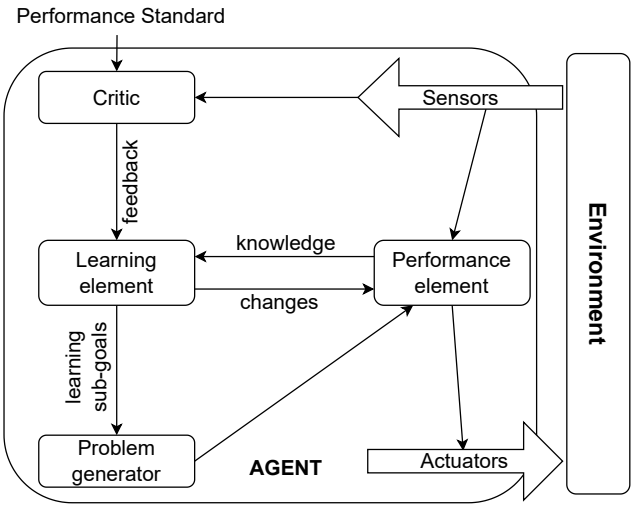

Fig. 2 Critic Agent Architecture (Russell and Norvig, 2021). We introduce this architecture as a baseline AI architecture that manages to capture various aspects of perceiving, learning, planning, reasoning and acting as different qualitative processes. One can visually contrast this architecture with with the other mainstream architectures, old and new, in AI.

humans in quite a few of these tasks (LeCun et al, 2015). However, like any supervised learning model, ANN's over-reliance on historical data means that they learn to repeat *what has been done*, not *what ought to be done*. Coupled with their largely black-box nature, this leads to a propagation of existing systemic biases that is difficult to identify or address. Post-hoc methods to interpret and 'explain' ANN-based models, such as LIME and SHAP (Samek et al, 2021), result not in an explanation of the internal mechanics of ANNs, but approximations in the form of equivalent interpretable models. For instance, in order to explain a deep-ANN, a decision tree or a heat-map are generated as an approximate function between the inputs and the outputs of the deep-ANN. This may, perhaps, be seen as a form of external, open loop reflection (but typically by others, not by the system itself); in and of themselves, the architecture of (feed-forward) ANNs has no capability for reflection.

Generative Adversarial Network Architectures (GAN)

GANs (Goodfellow et al, 2020) are one recent example of how ANNs can be used as building blocks within an explicitly designed architecture. These pitch two multi-layered perceptrons (ANNs) against each other in a 2-player mini-max game. The higher-level architecture here captures the sort of competitive creative co-adaptation found within co-evolutionary systems.

When it comes to the human ability of reflection, GANs by themselves are incapable of representing the process. While their architecture contains a feedback loop, it does not operate at the meta level: the architecture is 'flat'. It is not generally considered that a GAN (or coevolution in general) adds any type of high-level cognitive process. While ANNs in general are just

clusters of interconnected nodes with weighted edges, and the same may be said of the brain, we contend that there are essentially two approaches to generating cognitive processes of this type: one is a complex systems approach, where the virtual machine (Sloman and Chrisley, 2003; Sloman, 2013, 1996) operationalizing cognition emerges through complexity; the second is through an explicit architecture, as we do in this paper. Because AI agents that solely use ANNs such as these cannot reflect about themselves and the consequences of their actions in the world, they can behave anti-socially with no ability to know this.

One of the reflective components from Kolb's cycle (see Section 4, Fig. 1) missing here is Active Experimentation, which is distinct from exploration. Active Experimentation is also a multifaceted process, that includes at least exploration and active learning at the meta-level, and also intentional reconceptualizations of existing knowledge, i.e. Dennett's Tower of Generate-and-Test (Dennett, 1975, 2013), in order to be able to reflect on the value of new models.

Practical Reasoning Architectures

Procedural reflection provides a form of hard-coded first-order meta-reasoning the in Procedural Reasoning System (PRS), where a process is specified that deliberates over possible execution plans. The Procedural Reasoning System (based on Lisp (Smith, 1984) - see Fig. 4) implements this by passing symbols from the previous state (or lower symbolic level) to the current one such that we can say what the system was up to in the previous state (Smith, 1982).

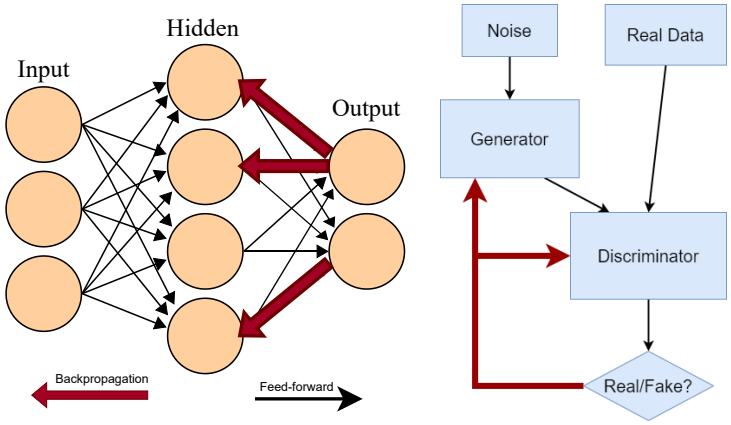

Fig. 3 ANN architecture (left, by Colin Burnett from https://en.wikipedia.org/wiki/Connectionism) and GAN (Goodfellow et al, 2020). Considering these common machine learning architectures, it is clear that there is a lack of any reflective 'loop'. Although these achieve different outcomes, they are qualitatively equivalent in the sense that they both operate at a single level of abstraction when it comes to information processing. There is no self-reference: the loops in both cases are for feedback, in much the same way that the Critic Agent operates. Additionally, even though Kolb's model of experiential learning is a model of learning in humans, it also presents (albeit at a high level) qualitative processes that ANNs and GANs do not.

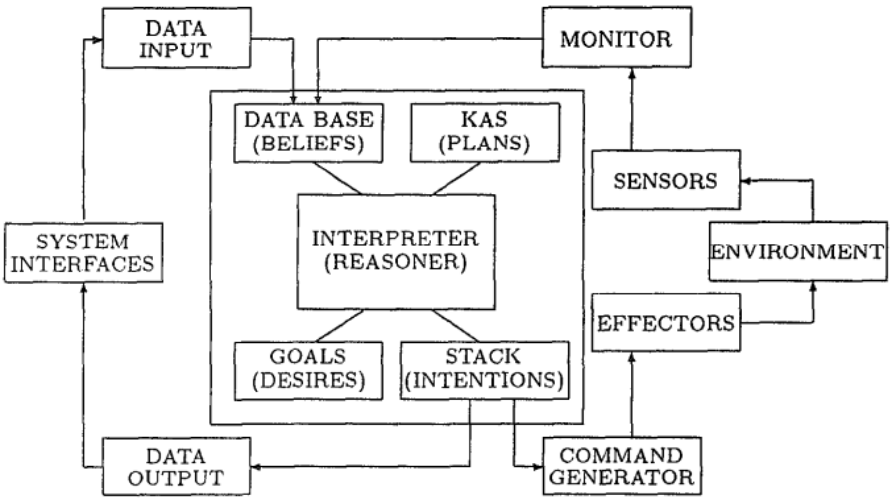

Fig. 4 Procedural Reasoning System (PRS) architecture (Source: (Georgeff and Lansky, 1987)). The PRS architecture is similar to the Critic agent architecture in the sense that it allows us to break down different qualitative processes. The difference between the Critic architecture and the PRS is that the PRS does not include a learning component, but it has a richer representation of the processes and elements responsible with driving the reasoning behind the actions that are executed in the environment. PRS also allows for an eventual learning component to be plugged into the system interface which feeds data into the belief base.

There is no recursive reasoning capability or ability to learn and integrate arbitrary self-models, both which are are crucial for many forms of reflection.

BDI Architectures

Later, Belief-Desire-Intention (BDI) architectures (Rao et al, 1995), Fig. 5, based on PRS, were introduced to structure cognitive reasoning based representations of and interactions between propositional models capturing the agent's beliefs, desires, and intentions.

Reflection has only recently been modelled in BDI architectures, again in the form of procedural reflection as in PRS. According to a recent survey on BDI agents (De Silva et al, 2020), only one paper was identified that implemented reflection in BDI architectures. In that work (Leask and Logan, 2018), BDI agents can use system-wide instructions to identify the context in which they operate and this enables them to use a rather reductive notion of reflection called procedural reflection to select deliberation strategies (Leask and Logan, 2018).

Both PRS and BDI architectures can efficiently allow for deliberation, which is distinct from our richer notion of reflection. Kolb's cycle includes the component of Abstract Conceptualisation (see Section 4, Fig. 1), which is missing from the process of procedural reflection. Without this component,

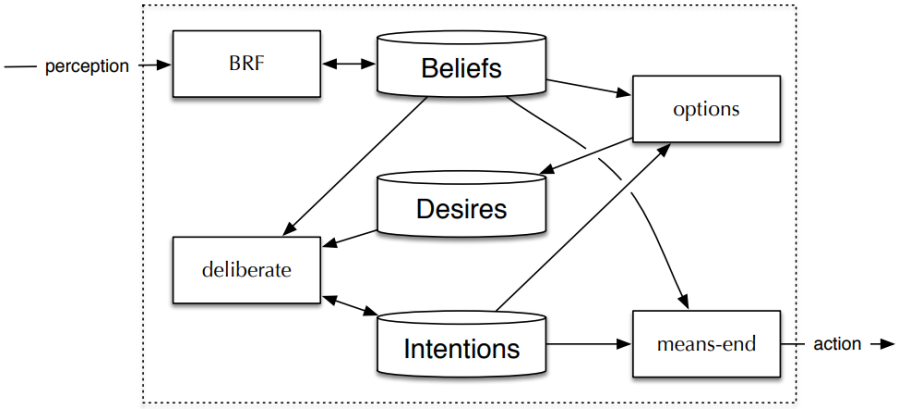

Fig. 5 The Belief-Desire-Intention (BDI) architecture (by Jomi F. Hübner from https://ai4industry.sciencesconf.org/data/Multi_Agent_Systems_lecture.pdf). The BDI architecture was designed to help AI agent designers build intuitive and interpretable AI agents capable of practical reasoning. The architecture depicts different qualitative processes and elements responsible for meta-reasoning (deliberation) and belief revision (BRF), which then help the agent decide what to do in a given circumstance in order to achieve their goals/desires in a dynamic environment.

deliberation is done without context. Regarding the difference between reflection and deliberation employed in practical reasoning, deliberation is a process for thinking out decisions, whereas reflection is a higher-level process that situates the agent that performs deliberation in a context through Abstract Conceptualisation. Deliberation does not require self-representation through Abstract Conceptualisation, because deliberation can be done at symbol-level, e.g., implementing deliberation strategies and selecting them using a procedural reflection.

Domain Expert Systems

These, such as tutoring expert systems, do not replicate reflection beyond in a rudimentary sense either. For advanced domain expert tutoring systems that are based on architectures like ACT-R (Anderson et al, 1997) and that implement some learning theory, they are reflective, but in the procedural sense that we explained above - Lisp style (see Fig. 6) (Smith, 1982). Another issue with systems like ACT-R is that they are architectures for domain expert systems where the environment is part of the system, not agent-based architectures like the critic agent architecture (Fig. 2) where agents act in an observed environment.

To summarise, the ANN architectures discussed above do not allow for reflection to be captured. Conversely, PRS, BDI, and ACT-R do not exclude it; neither do they explicitly describe it.

6 Building Reflective AI Agents

In order to make an agent reflective, thus expanding the list of Boden's 'sorts of things', we first need an architecture. We must separate out reflection from decision making and action.

Second, we need a suite of reflective cognition processes that may be included depending on the form of reflection desired. A given instance of a reflective agent may have one or more or all of these processes, in line with the differentiated theory approach. We categorise these processes in four tiers:

Tier 1 Reflective Agent: This incorporates models of self and others, and a process to reason using these models in order to ask itself what-if questions concerning generated actions. This enables a Popperian-style consequence engine and reflective governance process, able to evaluate proposed actions in context (acknowledging that context can change) and at least block some actions.

Tier 2 Reflective Agent: Adds processes that learn new reflective models, including incorporating feedback from new experiences into them incrementally. This addition enables Kolb-style reflective experiential learning.

Tier 3 Reflective Agent: Adds a reflective reasoning process that proposes not only a single 'optimal' solution, but is ready to present a diversity of possible ways forward – hypotheses to be tested – based on different approaches to solving the problem (including safe ways of disengaging from it).

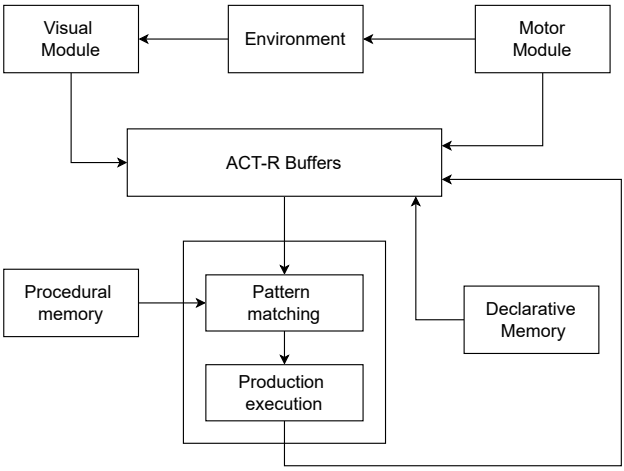

Fig. 6 ACT-R architecture (Anderson et al, 1997). It is crucial to note that ACT-R is not an AI agent architecture, rather a cognitive architecture that was used as an expert system. The original purpose of ACT-R is to map and understand human cognition as a set of modular components that execute procedures to produce behaviour in a specific domain. ACT-R assumes that all cognitive components are represented and driven by declarative and procedural memory.

There are many ways an agent may generate proposed actions. These vary in complexity substantially, from simple randomised search (e.g. mutation or exploration) through to heuristic and guided search approaches, up to potentially advanced forms of artificial creativity and strategic planning.

These approaches provide the ability to deliberate about novel possible strategies for action, including in new or potential (imagined) situations, and to evaluate these internally by reasoning with the reflective models.

Tier 4 Reflective Agent: Adds the ability to re-represent existing learnt models in new ways. This facilitates new reasoning possibilities and the potential for new insights. It provides a Gregorian-style ability to change the way the agent reflects.

Third, we need a way of representing the broader context: we need models of prevailing norms, and of the social values associated with the outcomes of different possible actions, and of other higher-level goals that may not be immediately or obviously relevant to the task.

Note that these components are mostly not new, but it is their novel combination and integration that provides new capability. Indeed, there are now several decades of work on reflective architectures, including early work like Landauer and Bellman's Wrappings (Landauer and Bellman, 1998), and Brazier and Treur's (Brazier and Treur, 1995) specification for agents that can reason reflectively about information states. More recently, Blum et al's Consequence Engine architecture (Blum et al, 2018), the EPiCS architecture (Lewis et al, 2015a) and the LRA-M architecture (Kounev et al, 2017b), are all aimed explicitly at achieving computational self-awareness through reflection.

On this broader point, self-awareness, often considered as the capacity to be 'the object of one's own attention' (Morin, 2006), has long been targeted as a valuable property for computational systems to possess (McCarthy, 1999; Mitchell, 2005), owing to the value of its functional role and evolutionary advantage in biological organisms (Lage et al, 2022). Computational forms of self-awareness require reflective processes that access, build, and operate on self-knowledge (Lewis et al, 2011; Kounev et al, 2017b). This self-knowledge is typically described according to five 'levels of self-awareness' (Lewis et al, 2011, 2015a, 2016) rooted in the work of Neisser (1997), although may consider many other aspects (Lewis et al, 2017). In some cases these are trivial self-models, for example a smartphone may have an internal parameter that captures whether its charging port contains moisture. Slightly more complex, the device may learn an internal model of its typical charging behaviour, sufficiently to act meaningfully on, and this may adapt as the battery degrades. In more complex examples still, a cyber-physical system may have a model of available resources discovered at run-time (Bellman et al, 2020).

Learning and reasoning with self-knowledge requires a reflective self-modelling process (Landauer and Bellman, 2016; Bellman et al, 2017) of the type described here. The exact form of such learning and self-modelling will vary depending on requirements and situation, but some examples include self-modelling based on abstraction from run-time data (e.g., Bellman et al

(2017)), or simulation of oneself in the environment (Blum et al, 2018; Elhabbash et al, 2021). As Blum et al demonstrate, such simulations may be used as 'consequence engines', similarly to how Hesslow (2002) describes the ability of the human brain to execute processes of internal cognitive simulation.

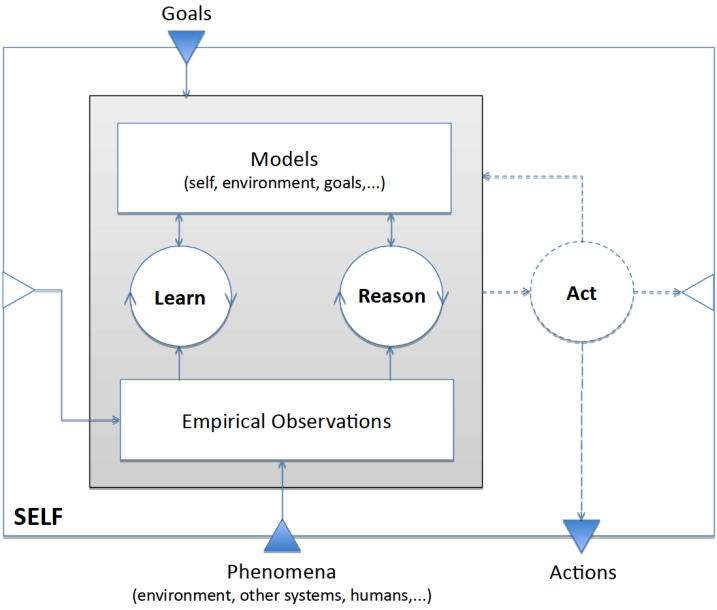

Fig. 7 LRA-M Reflective Architecture. Source: (Kounev et al, 2017b). The Learn-Reason-Act-Model (LRA-M) model was designed as a reference architecture to capture the essential components of computational self-awareness and their relations. Strictly speaking, acting is considered optional, depicted by the dashed line, though is typically the purpose of the self-awareness in a practical system. The circular arrows signify that learning and reasoning are ongoing processes at run-time, based on streaming data from ongoing observations of the world and oneself. Learning and reasoning also operate on existing internal models, including processes such as re-representation, abstraction, and planning.

The LRA-M model proposed by Kounev et al (2017b) (Figure 7), a commonly used reflective architecture that comes from the area of self-adaptive systems research, captures computational reflection at an abstract level. However, this leaves unclear several aspects associated with agents – e.g., what process generates the actions? Comparing this with a standard learning-based Critic agent (Russell and Norvig, 2021), we can see the inverse is true: learning and action selection are present, but reflection is not (see Figure 2).

Hence, here we propose one way to integrate the architecture of learning agents with the reflective schema captured by Kounev et al. In this way, a reflective architecture enables information to be abstracted and reasoned with at the meta-level, feeding back to update goals for learning, and to regulate behaviour.

We motivate our choice to base our architecture on Russell & Norvig's Critic Agent (Russell and Norvig, 2021), and further for using Kounev et al's LRA-M reflective loop (Kounev et al, 2017b) for discussing reflection in AI, since they enjoy broad understanding and acceptance in the domains of agent architecture and computational reflection, respectively. The critic agent, not because it is the best or most state-of-the-art for any particular domain, but because it allows us to illustrate how reflection can be incorporated into a very widely used and understood standard agent architecture. This, we hope, makes the article and argument more accessible. While there are also many reflective loops that we could have chosen, Kounev's is one that enjoys broad support, particularly from the self-adaptive systems community. Indeed, the article that presents that was the result of a large community effort at a Dagstuhl Seminar. Thus, while we acknowledge (and hope) that many other architectures can be paired with other forms of reflective loop, in this article, we use these two as an illustration and first step.

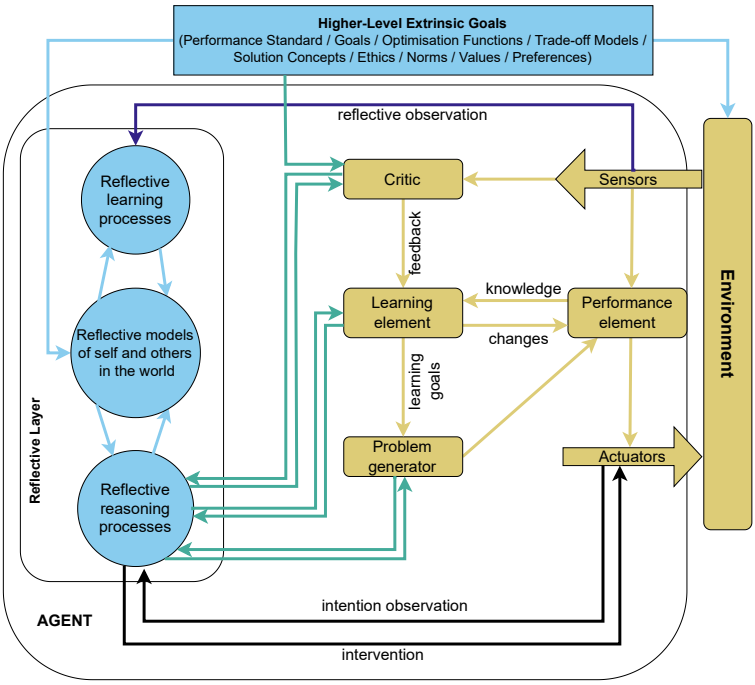

Fig. 8 Proposed Reflective Agent Architecture. The yellow elements in the centre and right columns are derived from the Critic Agent architecture Russell and Norvig (2021). Elements to the left and above, in blue, are added to form reflective capabilities and are derived from the LRA-M reflective architecture Kounev et al (2017b). Connections that integrate these, proposed by us, are depicted in purple and show processes of reflective observation and active experimentation (from Kolb (1984)) and behaviour governance (from Blum et al (2018)). Concrete elements are depicted in rounded rectangles. Circles are used in the reflective layer to indicate types of processes; these may be instantiated in various ways according to the form of reflection desired. An example instantiation is illustrated in Figure 9.

What we can now see is missing is a simple, generic schema for how reflection relates to existing agent architectures commonly used in modern AI. We propose such an architecture, as a synthesis of Russell and Norvig's Critic Agent and Kounev et al's LRA-M Architecture for reflective computational self-awareness, illustrated in Figure 8.

The advantage of an architectural approach is that it describes a separate set of processes, and we know that building systems that self-monitor is easier using an 'external layer' style (Weyns et al, 2013).

Indeed, while it would in principle be possible to propose an architecture that combines non-reflective and reflective cognitive activities in a single loop, this would be both unhelpful and misleading. First is that, as Weyns et al (2013) found, when designing agent architectures for self-reference, keeping these tiers of the architecture separate aids our ability to understand and analyze them. Second is that, as Sloman (2001) articulates, it is indeed the case that in cognitive systems there are several parallel processing 'loops' operating in parallel, and our architectures should make this explicit. In some cases, this may have the appearance of duplicating responsibilities, e.g., there are two arrows from 'Sensors' to different components, but in practice it simply means that there are different processes making use of the same information but in different ways. The choice of labelling these arrows in this architecture (e.g., 'Reflective Observation') therefore captures not only what information is passed between components, but what functionality that information passing enables as a part of the architecture.

Indeed, in early explorations of computational forms of reflection in software design (Maes, 1988), it was discovered that an architectural approach that factors out reflective (i.e. self-referential reasoning) processes from problem-focused reasoning processes enhances the elegance of the design. Note that what we propose is not an architecture that passes on information from one module to another as is the case in numerous hybrid approaches that aim to marry symbolic and sub-symbolic models (Calegari et al, 2020). What we propose is a cognitive architecture for reflection which can interpret information before passing it from one module to another. The interpretation of information is dynamic and happens in multiple processes. Below we describe how our proposed architecture ensures information interpretation at different cognitive levels through various 'reflective loops'.

There are indeed many reflective loops enabled by the addition of this new architectural capability. Here, we sketch some of the most obvious and perhaps important ones, particularly those that link to the conceptual discussion above. We categorised the loops according to the corresponding tiers of reflective agents:

Tier 1 Loop – Governance:

- Loop 1: Governing Behaviour:
  Actuators → Reflective Reasoning → Actuators

E.g. intervening to prevent an intended action.

Tier 2 Loops – Integrating experience and external factors:

- Loop 2: Abstract Conceptualization of Experience
  Sensors → Reflective Learning → Reflective Models → Reflective Reasoning → Critic.
  E.g. Kolbian experiential learning through conceptualization of new and changing experiences. Also calibrating and correcting exisitng models through new experiences.

- Loop 3: Learn about and integrate new extrinsic factors into operational goals:
  Higher-Level Extrinsic Goals → Environment → Reflecting Learning → Reflective Models → Reflective Reasoning → Critic.
  E.g. learning about and integrating new external factors, such as social norms, standards, and new user preferences, discovered in the environment, such as signs, verbal instructions, and observation of behaviour.

- Loop 4: Integrate new design goals into existing reflective models and operational goals:
  Higher-Level Extrinsic Goals → Reflective models → Reflective Reasoning → Critic.
  E.g. Directly integrating new goals, norms, preferences, and standards, that have been specified by an external operator rather than learnt from experience (as in Loop 3).

Tier 3 Loops – Critique and Imagination:

- Loop 5: Active Experimentation to Improve Potential Behaviour
  Actuators → Reflective Reasoning → Critic → Learning Element → Performance Element → Actuators
  E.g. Using the information that an action was intervened upon in order to adapt what the learning element learns, and hopefully avoid the situation in future. Or, creatively proposing novel courses of action and testing hypotheses regarding them.

- Loop 6: Reflecting on effectiveness of current operational goals and progress towards them:
  Reflective Reasoning → Critic → Reflective Reasoning.
  E.g. Counterfactual reasoning about current and potential goals, the 'what' of operational learning; black-box reasoning about progress towards them, for example asking 'am I stuck?' or 'would a different reward function better serve my high-level goals?'

- Loop 7: Reflecting on the current mechanisms of learning:
  Reflective Reasoning → Learning Element → Reflective Reasoning.
  E.g. White-box reasoning about current operational learning mechanisms, the 'how', for example asking 'how am I learning to do this?' and 'could I try to learn in a different way?'

Tier 4 Loop – Re-Representation:

- Loop 8: Reflective Thinking:
  Reflective Reasoning → Reflective Models → Reflective Reasoning
  E.g. refactoring models, finding and reconciling inconsistencies within and between models, determining areas for further hypothesis testing, re-representing existing conceptual knowledge in new formalisms and abstractions, concept synthesis.

One important and powerful insight is that these loops can be treated as 'primitives' and composed to provide additional and more complex cognitive features. For example, the composition of Loops 2 & 6 could give rise to curiosity-driven behaviour, while adding Loop 8 to this allows the result of the curiosity to be integrated into existing knowledge. Similarly, Loops 1 & 8 support reflecting on behaviour governance, for example reconciling competing imperatives, assessing the effectiveness of an intervention, or deliberating over an action. Adding Loop 6 to this, permits the deliberation to not only act over potential actions, but over potential directions for future learning.

Also of note is that there are now three arrows emanating from the 'Higher-Level Extrinsic Goals' box. This is since there are different ways in which these may reach and be acted on by a reflective agent (or not). First is that goals may be 'hard-coded' into the critic, as in Russell and Norvig's original model; in this case, an agent may have an intrinsic goal placed there by a designer, but the agent may not be aware of it in the reflective sense. Second, a goal may be 'hard coded' into the reflective layer as a goal model. Having this form of goal-awareness permits the reflective reasoner to integrate and contrast multiple goals, should that need arise, for example reconciling it with other goals and/or dismissing it when necessary, and to 'instruct' the critic to follow (combinations of) goals as necessary. Third, extrinsic goals may be communicated through the environment, being observed by sensors (e.g., noticing signage that communicates a desired behaviour in a new context, or being instructed by a human). In this case, the agent may need to identify, conceptualize, and integrate the goal, or decide that a goal communicated through the environment is counter to its existing goals, and need to engage in a process of reconciliation or other goal-based reasoning - i.e. a form of multi-agent deliberation (McBurney and Parsons, 2009; Tolchinsky et al, 2012) or practical reasoning (Atkinson and Bench-Capon, 2007). Taken together, these goal adoption mechanisms provide for a rich and multi-faceted way for goals to be

integrated by designers and as needs arise from the environment and interactions, and for the reflective layer to learn and reason about goals in complex ways. In this way, the architecture presents steps towards operationalizing the sort of nuanced goal-awareness proposed by Faniyi et al (2014), Lewis et al (2015b), and others.

There are a number of research challenges here. Specifically, there is a need to understand how to operationalize the above reflective loops, including operational semantics, APIs, and methods for the semantic transformation of information from symbolic to sub-symbolic levels and vice-versa.

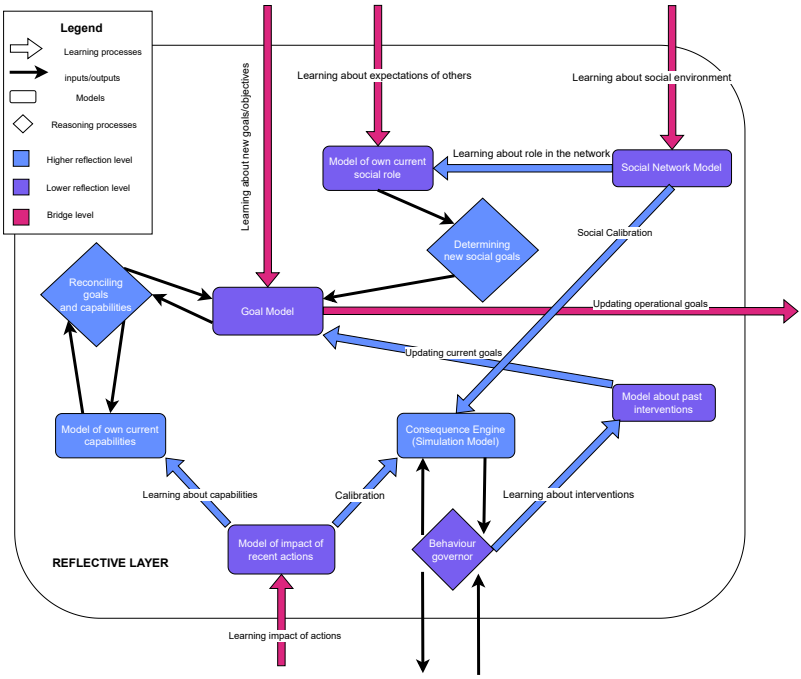

Fig. 9 Reflective Layer. An example instantiation of a possible reflective layer configuration. In this reflective layer, models and reasoning processes interact through passing information directly (input/output connections) and/or through learning processes. The level of reflection indicates that there is a cognitive hierarchy, where some processes and models operate at a higher level of reflection than others.

Similarly, the description of the above loops makes it clear that each of the reflective processes can be instantiated in various ways, depending on the input, output, and representation. The reflective layer in Figure 8 merely makes clear that reflective learning, reasoning, and models can exist, but does not provide detail of their instantiation. In general, we expect there to be a plethora of instantiations designed that adhere to this basic architectural pattern.

In Figure 9, we depict a possible instantiation of the reflective layer from Figure 8, as an illustration of this idea. Here, the Reflective layer has three

levels, namely the Bridge level, the Lower reflection level, and the Higher reflection level. It also has different components which are either situated at one of the levels, or pass information between themselves. Information between levels and modules is either passed directly, through input/output relations, or through learning and updating processes.

The Bridge level represents the processes which transform information such that the Lower reflection level is able to 1) use Active Experimentation to interact with the modules from Figure 8, such as the Critic, Learning element, Problem generator; 2) Learn about Higher-Level Extrinsic Goals; and 3) use Reflective Observation to interpret incoming information from the Sensors module.

The Lower level of reflection can also interact directly with the modules from 8 by skipping the Bridge level to use inputs and outputs in order to control the Actuator. This direct interaction with the non-reflective modules is done through the Behaviour governor.

The Lower level of reflection also interacts directly through inputs and outputs with the Higher level of reflection. For instance, it sends information to the reasoning process responsible for Determining new social goals

Active Experimentation can also improve the performance of the agent, by preparing it to deal with operational conflicts before the agent actually finds itself in the state of operational conflict.

Note that we are not proposing that any reflective agent must follow this instantiation; instead it is an illustration of how reflective learning and reasoning processes, along with the associated self-models, might be instantiated in a reflective layer.

7 How Do We Get There?

Many of the individual components required to realize Reflective AI already exist. In some cases, the challenge is to integrate these in a purposeful way to achieve the vision set out above. In other cases, there remain important fundamental research challenges. In this Section, we outline some of these in key areas.

7.1 Reflective Learning

Reflective Learning lies conceptually at the core of the proposed architecture. Fundamentally, learning here provides two forms of modelling capabilities: abstract conceptualisation and simulation, which support reasoning in complementary ways.

Abstract conceptualisation can be described as making sense of observations to form new ideas and theories (Kolb, 1984). More formally, this includes concept generation in the form of new classification schemes and formal models that represent new theories and conclusions about events that have been observed. This provides the agent with an interpretation of the experience at a different level of abstraction than the observations themselves. Done

computationally, this would enable agents with reasoning processes to make comparisons between competing understandings of a concept, comparing them against future empirical observations and upgrading them through ongoing adaptation. Further, by re-representing these models in new forms, the value of the model to the reasoning process may be similarly upgraded. For example, re-representing a simulation model in closed-form might enable more precise predictions, while re-representing an equation-based model in simulation might extend its predictive scope to capture the outcome of arbitrary or less-well-understood behaviours (e.g., Powers et al (2018)). Thus, through abstract conceptualisation and reflective reasoning over these models, the agent has a mechanism for hypothesis generation and testing for the purpose of both future action and cognition.

Note that this process of concept learning and re-representation is distinct from techniques that pass information between cognitive modules operating at the same level of abstraction (cf. Goodfellow et al (2020); Potter and De Jong (1994)). If the agent does not have abstraction capabilities, then the interaction more resembles a ping-pong game, where the outcome is each player improving in skill against the other, while the ball and rules retain the same form. An example of this might be a Generative Adversarial Network (GAN) (Goodfellow et al, 2020) producing ever-better fake faces, but never developing a model to theorise about properties of those faces. However, if these processes can transform the information to new abstractions, then instead a dialectic exists where new understanding can emerge.

However, Abstract Conceptualisation capabilities have not entered mainstream AI research yet. As illustrated in Figure 8, such a reflective process could start from Reflective Observation, which takes the data output of the Sensors and passes them to the Reflective Layer, where it uses a Reflective Learning process to transform this data into concepts that can then populate various new or updated self-models. Reasoning over these models can lead to intentional Active Experimentation, targeted at generating new experiences to observe, thus continuing the cycle.

Simulation models support a further form of reasoning, over consequences (Blum et al, 2018). This permits Dennett's 'Popperian' mind (Dennett, 2008), where hypothesis generation and testing can be carried out internally to the cognition of the agent, without requiring the world. For example, in the style of Hesslow (2002), an agent may build a simulation model in the form of a digital twin of itself in its environment. With sufficient interpretability and accompanied by automated reasoning processes, this may be complemented with an Abstract Conceptualisation, for example, that provides the understanding that the simulation contains an evolutionary stable strategy.

Note that neither the architecture nor the concept of reflective learning prescribe a particular learning algorithm. Many learning techniques can be used. The choice of technique itself is open-ended and can be made to suit the context so long as it adheres to, we posit, two conditions. First, that it is model-based, such that the process of learning produces a model that

captures some knowledge about the system and its environment. Interpretable models should be favoured, as the reasoning processes may then operate on these interpretations automatically. Second, that it operates online, such that it can incrementally build and update these models and be used in an anytime fashion.

Indeed, Lewis et al (2016) note that online and lifelong (Savage, 2022) learning algorithms are one of the key ingredients in achieving computational self-awareness. They further note that such online learning must be able to deal with concept drift, since both the system and its environment change. Wang et al (2016) show how existing online learning algorithms can be used for reflective self-awareness at different levels, but perhaps most importantly, they intentionally do not propose a preferred online learning paradigm, rather highlighting that empirical results suggest that using different learning techniques according to context can lead to enhanced performance. Complementary examples are presented in a collection edited by Pitt (2014), who arrives at a similar conclusion.

In the future, given a mechanism for representing concepts (Lieto, 2021), an AI agent could use Kolbian Abstract Conceptualisation (Kolb, 1984) to form new concepts and more meaningful models of itself and others in a shared system. Simultaneously, an agent could build simulation models of itself in its environment, to enable Popperian hypothesis testing. Both model forms provide complementary benefits (Powers et al, 2018) as forms of reflective modelling for meta-reasoning (Brazier and Treur, 1999), and in different ways, require the ability to learn models on-the-fly (Olteţeanu et al, 2019).

Research challenge: There is a need to develop mechanisms that learn human- and machine-interpretable conceptual and simulation models from empirical data and semantic information in the world, and further, to develop (unsupervised) methods for this to be done on the fly in a complex environment.

7.2 Reflective Governance

The proposed architecture captures Socrates's daemon (see Section 3 above) through a Blum-style governor loop (Blum et al, 2018) (also see Section 6), as mediator between Reflective Reasoning and an agent's Actuators. This loop is a process of deliberation at the meta-level. Reflection captures this process and situates it in a context, i.e., in an agent's model of the self and others in the world through Abstract Conceptualisation or simulation. Thus, the system does not need to re-learn its decision model if something in the set of oughts (Higher-level Extrinsic Goals) in its situation changes – though it might want to, later. It just needs to check the behaviour against them, and occasionally say 'no, that's not appropriate; give me an alternative, try a different approach.'

Regarding the ethical nature of this, explicitly ethical agents are nothing new, at least since Moor (2009) proposed a way of discerning four different 'types'. Indeed, the question of imbuing artificial agents with ethical values

was the topic of a special issue of Proceedings of the IEEE (Winfield et al, 2019). Both this and Cervantes et al's survey (Cervantes et al, 2020) provide an introduction. And indeed, Winfield and colleagues provided an early example (Blum et al, 2018) of putting these kinds of 'ethical governors' into robots, as consequence engines (Winfield and Jirotka, 2018); concerns also exist about whether explicit ethical agents are a good idea (Vanderelst and Winfield, 2018).

Research challenge: There is a need to develop inclusive, participatory methods for capturing values, norms, and preferences in formal, interpretable models that can be translated for use a) in a critic module to drive learning, b) as part of the behaviour governance process, and c) that respects the diversity of interpretation of human values that exists. There is a further need to develop governance and learning processes that adopt these in order to generate and ensure behaviour is aligned with them, as emphasised throughout the Royal Society Special Issue edited by Cath (2018).

7.3 Reflective Deliberation

Going deeper still, agents could extend the above with reflective deliberation. Reflective agents can deliberate by using Active Experimentation between Reflective Reasoning and Critic (see Figure 8) from time to time to find alternative ways of approaching problems. When considering finding multiple possible diverse and viable courses of action, we can draw on the rich and active research activity on dialogues, practical reasoning and value-based argumentation (Atkinson et al, 2005; Atkinson and Bench-Capon, 2007, 2016, 2021). These could help us to find new, different solutions, that come at a problem from a novel angle. And when evaluating these alternatives, we may choose to formulate the very notion of what 'successful' means according to our values; and in adopting these we must acknowledge that the best action may be a compromise.

Active Experimentation can drive deliberation to explore consequences an alternative courses of action which have not yet been instantiated. Additionally, this capability could be used to re-represent the reflective models of the world from different perspectives, including the agent's own goals, preferences for acting in certain ways, and analysing outcomes of internal simulations from a practical sense. The role of Active Experimentation here is also to improve deliberation on-the-fly, e.g. learning about an unforeseen consequence and integrating the knowledge about this consequence in a future deliberative process. This capability is crucial when dealing with potential operational conflicts, as it prepares the agent to deal with a situation which it has not encountered before in practice. Indeed, such processes have been developed and implemented in critical safety systems Bellman et al (2014), yet they are completely missing in areas which impact day-to-day lives, e.g. medical domain, or even human-AI interaction (see our Alexa 'smart' voice assistant example in Section 3). Current mainstream AI systems have neither deliberative reasoning capabilities nor the required internally simulated 'safe space' to perform Active Experimentation, i.e. a reflective model of the world where it simulates the

safety, ethical, moral, social etc. consequences of the actions suggested by their internal deliberative process.

To instantiate this sort of reflection, agents could employ value-based practical reasoning mechanisms such as action based alternating transition systems with values (AATS+V) or dialogue protocols (Sklar et al, 2013) together with an internally simulated domain. In turn, these are used to build argument schemes (Walton et al, 2008) which agents can use for both reflecting on their possible decisions (Sarkadi et al, 2019), as well as justifying their decisions by providing explanations (Mosca et al, 2020; Mosca and Such, 2021).

Research challenge: Agents need to be able to perform internal simulations of their actions and check the outcomes of these actions inside their own mind in order to perform deliberation. There is therefore a need to develop semantics and nested abstract models of the world for agent architectures to enable agents to go beyond the procedural reflection of BDI and PRS-like systems, by having the capability to run, analyze, and interpret new simulation models on the fly, according to need. One idea could be to develop polymorphic simulation models, that can be instantiated into specific simulations based on the learnt concepts and the need.

7.4 Social Context

Mentalistic capabilities, as we have explained in the chess example, play an important role in reflecting about one's complex decisions. Again, BDI-like agents can be given both the ability to communicate their decisions to other agents as well as the ability to model the minds of other agents inside their own cognitive architecture in order to better coordinate, or even delegate tasks (Rao et al, 1995; Sarkadi et al, 2018). Social interactions can be modelled and implemented with dialogue frameworks so that agents can explain and justify their behaviour (McBurney and Luck, 2007; Dennis and Oren, 2021). Modelling social context is a rich research field. Formal models of norms can be captured using deontic logic; research in normative systems considers the capturing of norms in agents (Criado et al, 2011) and human-robot interactions (Cranefield and Savarimuthu, 2021). Social context also includes social values represented in Higher-Level Extrinsic Goals. Solution Concepts (Ficici, 2004) give us one way to formalise these. These can be directly learned at the Reflective Layer by the agent through Reflective Learning. An AI system able to reflect on its actions in terms of social context would need to draw on formal models such as these. Work on agent-agent interoperability (Sarkadi et al, 2022; Sarkadi and Gandon, 2023), as well as work on normative reasoning in open MAS could play a crucial role, ranging from negotiation between individuals to engineering electronic institutions (Sierra et al, 1997, 2004; Pitt et al, 2012).

Research challenge: There is a need to develop the semantics and nested abstract models to refine the approaches described in (Criado et al, 2010; Criado, 2013; Sarkadi et al, 2018; Dennis and Oren, 2021), by integrating

socio-cognitive, communication and normative components inside the instantiated internal simulations. Reflective agents should be able to also simulate the minds and behaviours of other agents and organisations in various contexts where different norms are active, similarly to how Winfield's robots use it to predict the actions of other agents and anticipate the likely consequences of those actions both for themselves and the other agents (Winfield, 2018).

7.5 Implementation

A principal process of reflection is that of self-modelling for the purpose of reasoning about the changes one's actions brings about in an environment. This process still remains to this day very difficult to implement in machines that operate in complex environments (Nelson et al, 2022). As demonstrated by Nelson et al, one might start by looking at Fault Management Systems implementations for space operations. Another starting point would be to use test-beds for simulating cyber-physical systems. A good example of such a test-bed is CARS, which is based on the Wrappings software (Landauer, 2013). Work on implementing or engineering self-awareness has been mostly done in the area of autonomic computing (Kounev et al, 2017a). For instance, the SeAC Workshop series, which started with a Dagstuhl seminar in 2015 has been a dedicated forum to address this issue - see https://www.dagstuhl.de/15041. Another forum has been the AWARE workshop Cabri et al (2013). However, research challenges regarding self-awareness and self-modelling are different in AI systems, where the complexity of systems is entangled with a lack of system transparency. AI systems are themselves complex, adaptive and often opaque, compared with traditional computing systems, and therefore this represents an open challenge.

Regarding the metrics for evaluating implementations of reflective AI systems, one could follow a Distributed Processing Units approach, as described by Mertzani and Pitt (2022). Such an approach accounts for multiple metrics to be used in tandem as a cybernetic multi-agent system. This is useful for considering the social component of complex systems, going from the cyber-physical to the cyber-social. Another aspect that should be considered are the foundational properties of reflection within evolutionary and adaptive interactions of reflective AI agents. Aishwaryaprajna and Lewis (2023) do this by studying a co-evolving deliberative loop with neuroevolution that asks the agents to act with greedy or moderate behaviour in a sustainable foraging problem scenario. The novelty in Aishwaryaprajna and Lewis's work was adding a simple reflective governor to a neuroevolutionary controller, the latter being standard practice, but this being an example of how a reflective loop can be used to moderate the agents' behaviour and achieve sustainable behaviour and outcomes. While this was, we might argue, a very simplistic instance of an implementation of a reflective governor, it does show how part of the pattern described here could be effectively implemented using a modern neural controller.

Research challenge: There is a need to advance platforms and tools for making the modelling of complex systems easier and more intuitive. Another

challenge is defining metrics to find the right balance between complexity and efficiency, by integrating digital twins into self-* multi-agent systems (Tao and Qi, 2019). Designers and testers of reflective agents and systems should be able to use generic platforms and tools similar to those such as Tensorflow for deep learning (Abadi et al, 2016) and OpenAI Gym for evaluating reinforcement learning solutions (Brockman et al, 2016). Successfully addressing these challenges could unlock access for both researchers and practitioners in a range of diverse domains to be able to harness and build on these ideas.

8 Conclusion

Much research in AI is concerned with breaking a problem down until its constituent parts are solvable; this is important work. Conversely, linking these things together again in an agent-centric fashion to create the sorts of complex mind-like phenomena that motivated us in the first place, is just as crucial. As we have sketched above, there is a lot to draw on in conceiving and building reflective AI systems. Yet a lot of research remains in understanding how to put together the pieces of the puzzle. Some aspects of reflection are present in the established agent architectures and argumentation models for normative reasoning, deliberation, practical reasoning, and communication. After all, reflection is a crucial component of social interaction, cooperation, and reasoning about what others know and how they might act in different circumstances.

Returning to Weinberg (1972), the idea of Reflective AI is not about providing only scientific answers without any consideration of the broader socio-technical context. Reflective AI will be no silver bullet to the problems raised at the beginning of this paper, as they are fundamentally trans-scientific in nature. As such, it presents no excuse to avoid doing AI responsibly, and this would mean falling into the trap of what Oelschlaeger (1979) called 'the myth of the technological fix'. Delegating reflective mental capabilities does not nor cannot obviate human responsibility, nor should it distract from it. For example, when building and deploying AI systems, sadly too little attention is still often paid to making them context-sensitive, to understanding stakeholders and operational conditions, to requirements analysis, to understanding bias in data and how it might be amplified, to transparency about training sets, and to interpretability. What we are proposing here is not an either-or.

Instead what we are proposing is a socio-technical mechanism for providing social solutions to social problems, in the context of AI agent technology. To use an analogy, libraries are simply buildings, paper, and databases, that are built by and run by people, and enable us to enlighten, inform, and provide pleasure to the population at large. Reflective agents could be a set of methods, tools, and technologies that enable us to contextualise, socialise, put sensitivity into, enrich, and build trust with AI technology. In doing so, this agenda

aims to present a step towards a more complete, less unbalanced conceptualisation of AI systems that allows for more deliberate, mindful, and trustworthy technology, if we want to take it.

Acknowledgments

Ș.S. was supported by the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK Intelligence Community Postdoctoral Research Fellowship program.

This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

References

Abadi M, Barham P, Chen J, et al (2016) {TensorFlow}: a system for {Large-Scale} machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16), pp 265–283

Aishwaryaprajna, Lewis PR (2023) Exploring intervention in co-evolving deliberative neuro-evolution with reflective governance for the sustainable foraging problem. In: Artificial Life Conference Proceedings 35, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., p 140

Anderson JR, Matessa M, Lebiere C (1997) ACT-R: A theory of higher level cognition and its relation to visual attention. Human–Computer Interaction 12(4):439–462

Atkinson K, Bench-Capon T (2007) Practical reasoning as presumptive argumentation using action based alternating transition systems. Artificial Intelligence 171(10-15):855–874

Atkinson K, Bench-Capon T (2016) States, goals and values: Revisiting practical reasoning. Argument & Computation 7(2-3):135–154

Atkinson K, Bench-Capon T (2021) Value-based argumentation. Journal of Applied Logics 8(6):1543–1588

Atkinson K, Bench-Capon T, McBurney P (2005) Multi-Agent Argumentation for eDemocracy. In: EUMAS, pp 35–46

Bellman K, Landauer C, Dutt N, et al (2020) Self-aware cyber-physical systems. ACM Transactions on Cyber-Physical Systems 4(4)

Bellman KL, Nelson PR, Landauer C (2014) Active experimentation and computational reflection for design and testing of cyber-physical systems. In: CSDM (Posters), Citeseer, pp 251–262

Bellman KL, Landauer C, Nelson P, et al (2017) Self-modeling and self-awareness. In: Kounev S, Kephart JO, Milenkoski A, et al (eds) Self-Aware Computing Systems. Springer, p 279–304

Bellman R (1978) An Introduction to Artificial Intelligence: Can Computers Think? Boyd & Fraser, San Francisco

Blum C, Winfield AF, Hafner VV (2018) Simulation-based internal models for safer robots. Frontiers in Robotics and AI 4:74

Boden MA (1998) Creativity and artificial intelligence. Artificial intelligence 103(1-2):347–356

Boden MA (2016) AI: Its Nature and Future. Oxford University Press

Brazier F, Treur J (1995) Formal specification of reflective agents. In: IJCAI '95 Workshop on Reflection, M. Ibrahim, ed. Montreal, pp 103–112

Brazier FM, Treur J (1999) Compositional modelling of reflective agents. International Journal of Human-Computer Studies 50(5):407–431

Brockman G, Cheung V, Pettersson L, et al (2016) Openai gym. arXiv preprint arXiv:160601540

Cabri G, Hart E, Pitt J (2013) 3rd aware workshop on challenges for achieving self-awareness in autonomic systems. In: 2013 IEEE 7th International Conference on Self-Adaptation and Self-Organizing Systems Workshops, IEEE, pp xv–xvi

Calegari R, Ciatto G, Omicini A (2020) On the integration of symbolic and sub-symbolic techniques for XAI: A survey. Intelligenza Artificiale 14(1):7–32

Cath C (2018) Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376(2133):20180,080

Cervantes JA, López S, Rodríguez LF, et al (2020) Artificial moral agents: A survey of the current status. Science and Engineering Ethics 26(2):501–532

Cranefield S, Savarimuthu BTR (2021) Normative multi-agent systems and human-robot interaction. In: Workshop on Robot Behavior Adaptation to Human Social Norms (TSAR), pp 1–3

Criado N (2013) Using norms to control open multi-agent systems. AI Communications 26(3):317–318

Criado N, Argente E, Botti V (2010) A bdi architecture for normative decision making. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1, pp 1383–1384

Criado N, Argente E, Botti V (2011) Open issues for normative multi-agent systems. AI communications 24(3):233–264

De Silva L, Meneguzzi F, Logan B (2020) Bdi agent architectures: A survey. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), 2020, Japão., International Joint Conferences on Artificial Intelligence

Dennett DC (1975) Why the law of effect will not go away. Journal for the Theory of Social Behaviour 5:169–187

Dennett DC (1996) Kinds of minds: Toward an understanding of consciousness. Basic Books

Dennett DC (2008) Kinds of minds: Toward an understanding of consciousness. Basic Books

Dennett DC (2013) The role of language in intelligence. Walter de Gruyter

Dennis LA, Oren N (2021) Explaining bdi agent behaviour through dialogue. In: Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS)

Dignum V, Dignum F (2020) Agents are dead. Long live agents! In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, pp 1701–1705

Elhabbash A, Bahsoon R, Tino P, et al (2021) Attaining meta-self-awareness through assessment of quality-of-knowledge. In: 2021 IEEE International Conference on Web Services (ICWS). IEEE Computer Society, pp 712–723

Faniyi F, Lewis PR, Bahsoon R, et al (2014) Architecting self-aware software systems. In: 2014 IEEE/IFIP Conference on Software Architecture, pp 91–94

Ficici SG (2004) Solution concepts in coevolutionary algorithms. PhD thesis, Brandeis University

Georgeff MP, Lansky AL (1987) Reactive reasoning and planning. In: AAAI, pp 677–682

Goodfellow I, Pouget-Abadie J, Mirza M, et al (2020) Generative adversarial networks. Communications of the ACM 63(11):139–144

Hesslow G (2002) Conscious thought as simulation of behaviour and perception. Trends in cognitive sciences 6(6):242–247

Hesslow G (2012) The current status of the simulation theory of cognition. Brain research 1428:71–79

Kolb DA (1984) Experiential learning: Experience as the source of learning and development. Prentice-Hall, Englewood Cliffs, N.J., USA

Kounev S, Kephart JO, Milenkoski A, et al (eds) (2017a) Self-Aware Computing Systems. Springer

Kounev S, Lewis P, Bellman K, et al (2017b) The notion of self-aware computing. In: Kounev S, Kephart JO, Milenkoski A, et al (eds) Self-Aware Computing Systems. Springer, p 3–16

Lage CA, Wolmarans DW, Mograbi DC (2022) An evolutionary view of self-awareness. Behavioural Processes 194:104,543

Landauer C (2013) Infrastructure for studying infrastructure. In: 2013 Workshop on Embedded Self-Organizing Systems (ESOS 13)

Landauer C, Bellman KL (1998) Wrappings for software development. In: Proceedings of the Thirty-First Hawaii International Conference on System Sciences, pp 420–429 vol.3

Landauer C, Bellman KL (2016) Reflective systems need models at run time. In: Götz S, Bencomo N, Bellman KL, et al (eds) Proceedings of the 11th International Workshop on Models@run.time co-located with 19th International Conference on Model Driven Engineering Languages and Systems (MODELS 2016), Saint Malo, France, October 4, 2016, CEUR Workshop Proceedings, vol 1742. CEUR-WS.org, pp 52–59, URL http://ceur-ws.org/Vol-1742/MRT16_paper_10.pdf

Leask S, Logan B (2018) Programming agent deliberation using procedural reflection. Fundamenta Informaticae 158(1-3):93–120

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

Lewis PR, Marsh S (2021) What is it like to trust a rock? a functionalist perspective on trust and trustworthiness in artificial intelligence. Cognitive Systems Research 72:33–49

Lewis PR, Chandra A, Parsons S, et al (2011) A Survey of Self-Awareness and Its Application in Computing Systems. In: Proceedings of the International Conference on Self-Adaptive and Self-Organizing Systems Workshops (SASOW). IEEE Computer Society, pp 102–107

Lewis PR, Chandra A, Faniyi F, et al (2015a) Architectural aspects of self-aware and self-expressive computing systems. IEEE Computer 48:62–70

Lewis PR, Chandra A, Faniyi F, et al (2015b) Architectural aspects of self-aware and self-expressive computing systems. IEEE Computer 48:62–70

Lewis PR, Platzner M, Rinner B, et al (eds) (2016) Self-Aware Computing Systems: An Engineering Approach. Springer

Lewis PR, Bellman KL, Landauer C, et al (2017) Towards a framework for the levels and aspects of self-aware computing systems. In: Kounev S, Kephart JO, Milenkoski A, et al (eds) Self-Aware Computing Systems. Springer, p 3–16

Lieto A (2021) Cognitive design for artificial minds. Routledge

Maes P (1988) Computational reflection. The Knowledge Engineering Review 3(1):1–19

Mayor A (2018) Gods and Robots: Myths, Machines, and Ancient Dreams of Technology. Princeton University Press

McBurney P, Luck M (2007) The agents are all busy doing stuff! IEEE Intelligent Systems 22(4):6–7

McBurney P, Parsons S (2009) Dialogue games for agent argumentation. Argumentation in artificial intelligence pp 261–280

McCarthy J (1999) Making robots conscious of their mental states. In: Machine Intelligence 15, Intelligent Agents [St. Catherine's College, Oxford, July 1995]. Oxford University, p 3–17

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5(4):115–133

Mertzani A, Pitt J (2022) Metrics for reflection in distributed information processing. In: Proc. 14th Int. Workshop Agent-Based Modelling Hum. Behav.(ABMHuB)

Mitchell M (2005) Self-awareness and control in decentralized systems. In: Metacognition in Computation. AAAI Spring Symposium, p 80–85

Monett D, Lewis CWP, Thórisson KR, et al (2020) Special issue "on defining artificial intelligence" – commentaries and author's response. Journal of Artificial General Intelligence 11:1–100

Moor JH (2009) Four kinds of ethical robots. Philosophy Now 72:12–14

Morin A (2006) Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. Consciousness and Cognition 15:358–71

Mosca F, Such J (2021) Elvira: an explainable agent for value and utility-driven multiuser privacy. In: International Conference on Autonomous Agents and Multiagent Systems (AAMAS)

Mosca F, Sarkadi Ş, Such JM, et al (2020) Agent EXPRI: Licence to explain. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer, pp 21–38

Neisser U (1997) The roots of self-knowledge: Perceiving self, it, and thou. Annals of the New York Academy of Science 818:19–33

Nelson PR, Bellman KL, Landauer C (2022) Self-modeling-a practical example of why it's hard. In: 2022 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), IEEE, pp 88–94

Nesselrath HG, Russell D, Cawkwell G, et al (eds) (2010) On the daimonion of Socrates: Plutarch. SAPERE, Mohr Siebeck GmbH and Co. KG

Oelschlaeger M (1979) The myth of the technological fix. The Southwestern Journal of Philosophy 10(1):43–53

Olteţeanu AM, Schöttner M, Bahety A (2019) Towards a multi-level exploration of human and computational re-representation in unified cognitive frameworks. Frontiers in psychology 10:940

Pitt J (ed) (2014) The Computer After Me. Imperial College Press / World Scientific

Pitt J, Schaumeier J, Artikis A (2012) Axiomatization of socio-economic principles for self-organizing institutions: Concepts, experiments and challenges. ACM Transactions on Autonomous and Adaptive Systems (TAAS) 7(4):1–39

Plato (translated by Paul Shorey) (1969) Plato in Twelve Volumes, vol 5 & 6. Harvard University Press, Cambridge, MA, USA

Potter MA, De Jong KA (1994) A cooperative coevolutionary approach to function optimization. In: International Conference on Parallel Problem Solving from Nature, Springer, pp 249–257

Powers ST, Ekárt A, Lewis PR (2018) Modelling enduring institutions: The complementarity of evolutionary and agent-based approaches. Cognitive

Systems Research 52:67–81

Powers ST, Linnyk O, Guckert M, et al (2023) The Stuff We Swim in: Regulation Alone Will Not Lead to Justifiable Trust in AI. IEEE Technology and Society Magazine 42(4):95–106

Rao AS, Georgeff MP, et al (1995) BDI agents: From theory to practice. In: ICMAS, pp 312–319

Reuters (2018) Amazon ditched AI recruiting tool that favored men for technical jobs. The Guardian Oct. 11, 2018

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 65(6):386

Russell S, Norvig P (2021) Artificial intelligence: A modern approach, global edition 4th. Foundations 19:23

Samek W, Montavon G, Lapuschkin S, et al (2021) Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE 109(3):247–278

Sarkadi S, Gandon F (2023) Interoperable AI for Self-Organisation. In: 2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), IEEE, pp 86–87

Sarkadi Ş, Panisson AR, Bordini RH, et al (2018) Towards an approach for modelling uncertain theory of mind in multi-agent systems. In: International Conference on Agreement Technologies, Springer, pp 3–17

Sarkadi Ş, Panisson AR, Bordini RH, et al (2019) Modelling deception using theory of mind in multi-agent systems. AI Communications 32(4):287–302

Sarkadi S, Tettamanzi AG, Gandon F (2022) Interoperable AI: Evolutionary Race Toward Sustainable Knowledge Sharing. IEEE Internet Computing 26(6):25–32

Savage N (2022) Learning over a lifetime. Nature

Schön DA (1984) The Reflective Practitioner: How Professionals Think In Action. Basic Books

Sierra C, Jennings NR, Noriega P, et al (1997) A framework for argumentation-based negotiation. In: International Workshop on Agent Theories, Architectures, and Languages, Springer, pp 177–192

Sierra C, Rodriguez-Aguilar JA, Noriega P, et al (2004) Engineering multi-agent systems as electronic institutions. European Journal for the Informatics Professional 4(4):33–39

Sklar EI, Azhar MQ, Parsons S, et al (2013) A case for argumentation to enable human-robot collaboration. Proceedings of Autonomous Agents and Multiagent Systems (AAMAS), St Paul, MN, USA

Sloman A (1996) What is it like to be a rock?, URL https://www.cs.bham.ac.uk/research/projects/cogaff/misc/rock/

Sloman A (2001) Varieties of affect and the CogAff architecture schema pp 39–48

Sloman A (2013) Virtual machine functionalism: The only form of functionalism worth taking seriously in philosophy of mind, URL https://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html

Sloman A, Chrisley R (2003) Virtual machines and consciousness. Journal of Consciousness Studies 10:133–172

Smith BC (1982) Procedural reflection in programming languages. PhD thesis, Massachusetts Institute of Technology

Smith BC (1984) Reflection and semantics in lisp. In: Proceedings of the 11th ACM SIGACT-SIGPLAN symposium on Principles of programming languages, pp 23–35

Sun R (2001) Cognitive science meets multi-agent systems: A prolegomenon. Philosophical psychology 14(1):5–28

Tao F, Qi Q (2019) Make more digital twins. Nature 573(7775):490–491

Tine M (2009) Uncovering a differentiated theory of mind in children with autism and asperger syndrome. PhD thesis, Boston College, USA

Tolchinsky P, Modgil S, Atkinson K, et al (2012) Deliberation dialogues for reasoning about safety critical actions. Autonomous Agents and Multi-Agent Systems 25(2):209–259

Vanderelst D, Winfield AFT (2018) The dark side of ethical robots. In: AAAI/ACM Conference on AI Ethics and Society, pp 317–322

Walton D, Reed C, Macagno F (2008) Argumentation schemes. Cambridge University Press

Wang S, Nebehay G, Esterle L, et al (2016) Common techniques for self-awareness and self-expression. In: Lewis PR, Platzner M, Rinner B, et al

(eds) Self-Aware Computing Systems: An Engineering Approach. Springer, p 113–142

Weinberg AM (1972) Science and trans-science. Science 177(4045):211–211

Weyns D, Iftikhar MU, Söderlund J (2013) Do external feedback loops improve the design of self-adaptive systems? a controlled experiment. In: 2013 8th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS), IEEE, pp 3–12

Winfield AF (2018) Experiments in artificial theory of mind: From safety to story-telling. Frontiers in Robotics and AI 5:75

Winfield AF, Jirotka M (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376(2133):20180,085

Winfield AF, Michael K, Pitt J, et al (2019) Machine ethics: The design and governance of ethical ai and autonomous systems [scanning the issue]. Proceedings of the IEEE 107(3):509–517