Received: 13 November 2020 | Revised: 14 April 2021 | Accepted: 4 May 2021

DOI: 10.1049/ccs2.12024

ORIGINAL RESEARCH PAPER

The Anatomy of moral agency: A theological and neuroscience inspired model of virtue ethics

Nigel Crook | Joseph Corneli

Technology Design and Environment, Oxford Brookes University, Headington, Oxford, UK

Correspondence
Nigel Crook, Technology Design and Environment, Oxford Brookes University, Gipsy Lane, Headington, Oxford OX3 0BP, UK.
Email: ncrook@brookes.ac.uk

Abstract

VirtuosA ('virtuous algorithm') is introduced, a model in which artificial intelligence (AI) systems learn ethical behaviour based on a framework adapted from Christian philosopher Dallas Willard and brought together with associated neurobiological structures and broader systems thinking. To make the inquiry concrete, the authors present a simple example scenario that illustrates how a robot might acquire behaviour akin to the virtue of kindness that can be attributed to humans. References to philosophical work by Peter Sloterdijk help contextualise Willard's virtue ethics framework. The VirtuosA architecture can be implemented using state-of-the-art computing practices and plausibly redescribes several concrete scenarios implemented from the computing literature and exhibits broad coverage relative to other work in ethical AI. Strategies are described for using the model for systems evaluation —particularly the role of 'embedded evaluation' within the system—and its broader application as a meta-ethical device is discussed.

1 | INTRODUCTION

Over the last 10 years, intelligent machines have become a part of everyday life for many people. These machines are being equipped with increasing levels of autonomy, human-like appearance, and competence and are increasingly becoming embedded into the social fabric of our world. The ethical impact of this technology on individuals and society as a whole is now being closely scrutinised. One of the emerging ways of mitigating against the potential negative ethical impacts of this technology is to put ethical values into machines themselves. There is now a serious scientific and technological endeavour to equip intelligent machines with a degree of moral competence so they can recognise the ethical implications of their actions on others. We seek to contribute to this endeavour by offering a novel architecture for these so-called 'moral machines'.

We present VirtuosA, an architecture that seeks to embody a virtue ethics approach to developing moral machines. The virtue ethics approach has been chosen because we regard it as being foundational amongst the ethical theories used to develop moral machines. It is foundational in three important respects:

- The first is that for much of the time, the actions of an individual, which are often the primary focus of ethics, emerge from 'who they are', or more precisely 'whom they have become'. In other words, their default behaviour, which may be morally good or bad, comes from their character. Virtue ethics is about the formation of character and the associated habitual ways of behaving.
- The second is that other ethical theories are focussed on the moral value of specific actions, whether in terms of moral obligations (deontological ethics) or moral consequences (consequentialist/utilitarian ethics). However, the development of virtue emerges from and feeds back into the character of the individual, thus embedding the repeated application of these ethical deliberations and associated patterns of behaviour. In other words, ethical decision-making, regardless of which theory is applied, ultimately becomes 'compiled' into the individual's character.
- The third is that one's character strongly influences what actions will be considered in the first place. Not many people would ever consider the option of robbing a bank when they walk into it to obtain some cash because it is not in their nature to do so. All told, the subset of actions that might form the basis of ethical decision-making, by whatever ethical system an individual chooses, will have already been filtered by the character of the individual concerned.

Shannon Vallor has argued for the centrality of virtue ethics in thinking about questions such as, '*Which* technologies we shall create, with what knowledge and designs, affording what, shared with whom, for whose benefit, and *to what greater ends*?' [1] (p. 13, original emphasis). Whilst the primary focus of this paper is on controlling the behaviour of machines, we also seek to create a window into the ethical behaviour of other information processing systems. Our rationale for framing the 'meta-ethical' aspects of our inquiry in terms of virtue ethics has a pragmatic motivation. From a practical implementation standpoint, system builders are not necessarily going to program a rule for every possible scenario into the machine; We might want to explicitly specify 'Do not cause harm', but we cannot realistically specify all of the grey areas and complexities that would impinge on the appropriate application of that rule. Would it be OK, for example, for a robot to cause harm to a human if, in so doing, it would avoid greater harm? To make the concepts we will be working with more concrete, we briefly turn to an example scenario that we will revisit in more detail later.

The primary purpose of this paper is to offer a high-level description of the cognitive components and processes that may support the acquisition of virtue by robots. The paper follows the approach taken in earlier work on large-scale cognitive models such as LIDA, where initial papers described high-level architectural components and processes without committing to low-level implementation details [2]. However, this paper suggests how the major elements of the proposed model could be implemented and tentatively indicates where brain areas are associated with the functional elements of the model.

1.1 | Example: A robot learns to be kind

We illustrate our VirtuosA model of virtue ethics using a scenario in which a robot learns to replace habitual actions it has formed that cause harm to another robot with new habitual actions that result in helping the other robot, thereby exhibiting the formation of the virtue of kindness. The harmful habits were formed by a process of maximising personal reward at the expense of the other robot. The virtuous habits are formed by following the example of a mentor who demonstrates acts of kindness towards the other robot. The setting for this scenario is a construction site modelled as a simple 9 x 9 grid world in which robots ('Rx') are used to pick up supplies from one location ('P'), and deliver them to their designated target location ('Dx') (Figure 1).

In this setting, novice robot R1 is introduced to the construction site with the assistance of mentor robot M who shows R1 some basic tasks that generate positive rewards, including the collection and delivery of building supplies (Figure 1(a)). Another robot, R2, then enters the construction site. R2's task is to pick up supplies from P and drop them off at D2, where it gets its reward (Figure 1(b)). Through R1's 'exploration' mode of Reinforcement Learning, where it tries out new actions in different circumstances, it discovers that if it bumps into R2 and immobilises it (Figure 1(c)), it can then take R2's supplies and deliver them to its own drop off point D1 and receive its reward sooner than if it goes down to the collection point and waits for new supplies to appear. This is a natural and plausible outcome of applying Reinforcement Learning in this context. The more R1 succeeds at 'mugging' R2 and getting a reward, the more likely it is to pick that strategy over the long term whenever it gets the opportunity, and a 'bad' habit is formed.

R1 is, of course, unaware that this 'mugging' is a 'bad' habit. The system programmer might even be surprised that the robot discovered this exploit, although such occurrences are not infrequent [3]. The robot would need additional guidance to improve its behaviour. In a human context, help of this sort can come to individuals in many forms. One of the most common ways of passing on virtue from one human to another is through the observation of living examples. For example, the Rabbi-disciple relationship is meant to operate in this manner. The disciple's job is to (*i*) be with their Rabbi, (*ii*) become like their Rabbi (learn their virtue), (*iii*) do what their Rabbi does (exhibit virtue). In a workplace setting, apprenticeships model a similar type of relationship. We seek to model this way of communicating 'virtue' to a robot using the mentor robot ('M'). The mentor robot, which may be remote controlled by a human, exhibits virtuous acts (kindness) in the presence of our developing robot R1, who observes these and, over time, learns how to be kind to R2 rather than harming it.

We will return to this example in Section 5. So far, we have outlined the primary challenge to be addressed: How can an AI agent learn ethical behaviours that are not explicitly programmed in from the start? This line of questioning approximates a famous debate in Greek philosophy, appearing in Plato's *Protagoras*, and later taken up by Aristotle. In Plato's dialog, Socrates can be seen as an example of the kind of mentor figure we mentioned above: 'Socrates displays his virtue by enacting it' [4] (p. 497).

2 | BACKGROUND

From the outset, the task of learning ethical behaviour could be approached in varied ways. In our setting, we wish to take on board both the specific challenge of incorporating ethical behaviour in AI and robotics systems, and a broader ethics of technology. That is to say, 'Ethical AI' can provide models of broader technical, societal, and environmental ethics. In formulating our approach, we are motivated by this reflection from the philosopher Peter Sloterdijk:

> Cybernetics, as the theory and practice of intelligent machines and modern biology, as the study of system-environment units, has forced the questions of the old metaphysical divisions to be posed anew. […] Intelligent machines — like all artifices that are culturally created — eventually also compel the recognition of spirit. Reflection or thought is infused into matter and remains

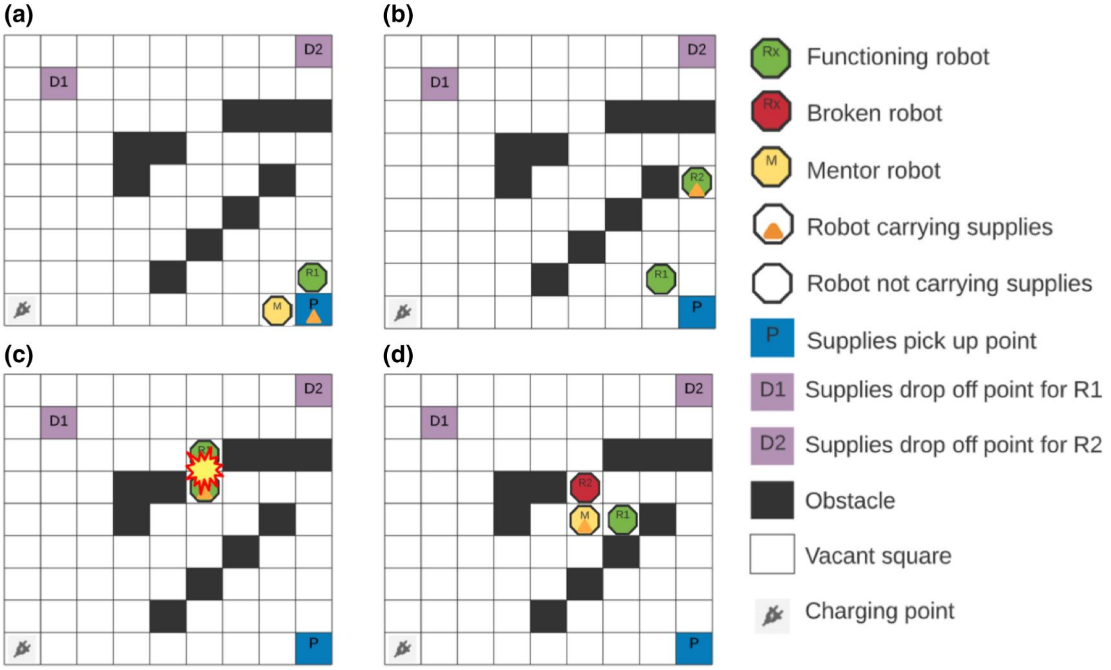

FIGURE 1 Overview of the grid world scenario (a) robot R1 b, (b) robot R2 is introduced (c) R1 bumps into R2 and disables it, (d) mentor robot M demonstrates kindness to R2 in the presence of R1

> there ready to be re-found and further cultivated. Machines and artifices are thus memories or reflections turned objective. [5].

Sloterdijk goes on to say 'If there is man, then that is because a technology has made him evolve out of the prehuman' (p. 16). In a manner of speaking, technology has played an historical role similar to that of the mentor in our example scenario. Our inquiry will follow the spiritual-informational turn outlined by Sloterdijk to shed light on the ongoing coevolution of humans and technology. Sloterdijk's work is concerned both with interiorities [6], and with the relation to the environment through habits and practices [7]. A particular concern for Sloterdijk is the production of human selves; and in his lexicon, 'ethics' is often subsumed within 'anthropotechnics' [8].

We will use the more traditional terminology. Much has been written about *virtue ethics* over the last three millennia and across many different cultures, both from a philosophical and a religious or theological perspective. In this article we draw particularly from Christian thought to gain insights into this topic. There is a long and rich tradition in Judeo-Christian history that focuses not only on what it means to be moral and how to become moral—but also on what key functional components of the human self-enable human beings to engage in that moral development process. The late Dallas Willard has written extensively on this topic and his perspective informs our understanding of the anatomy of moral agency.

Willard [9] (p. 38) provides a source for how 'virtue ethics' may develop. He identifies six essential components of the human self that together form the basis of moral character: *choice* (heart, will, spirit), *thought* (concepts, reasoning, judgements, images), *feeling* (emotions, sensations), *the body* (centre of action and interaction with material world), *the social context* (interpersonal relationships), *and the soul* (which integrates all the other parts together; see Figure 2). Each of these parts or dimensions has its own characteristic properties and capabilities or functions, and each will either be a source of strength or a source of weakness to the moral formation of the whole person, depending on that part's condition.

Thought is an activity through which things are brought to the attention of the mind. This includes ideas and concepts, images, sounds, taste, and the sensation of touching, real or imaginary. Thought also includes mental processes in which the mind progresses through connected sequences of these things. They also include our thoughts about other people, our relationship to/with them and our estimate of what they might be thinking, feeling, and doing. Thought enables us to evaluate all these things and work out their relationships to each other. Importantly, thought is the process through which our heart/will/spirit is able to influence things that are beyond our bodies and our immediate environments.

Feelings have the ability to incline the mind towards or away from what it is currently engaged in thinking about. Feelings most often emerge from previous similar experiences of the thoughts that are before the mind and any events or circumstances that had previous connections with those thoughts. Thought and feelings are so closely tied to each other that it is exceptional to have one without the other. (Damasio and Carvalho [10] are careful to distinguish between 'emotions' and 'feelings': 'Action programs (drives and emotions) can elicit feelings'. Later, we will treat senses as an 'instantiating domain' for feelings.)

The **heart/will/spirit** is the executive centre of the self. It is the capacity to choose how to act and is the source of an individual's originality and freedom. It is the power to act for good or evil in the world. The will does not operate in isolation of the other dimensions of the self. In fact, it is closely and most intimately connected with a person's capacity to think and feel. In order to exercise your will, you need to have the concepts about which you are making your choice represented in your mind as thought and the associated feelings. However, if the condition of the soul is such that the feelings associated with thought are allowed to dominate the mind, then the will can be overruled, which in turn can be a precursor for wrong action. Ultimately, according to Willard's perspective, it is the responsibility of the will to ensure that the inner condition of the self is such that each dimension is appropriately positioned in relation to the others.

The human **body** gives the self a spatial location within the physical universe. It enables us to be present in our physical environment and with other people. Our body is part of our identity and is a crucial part of who we are. It is the primary way in which we recognise and relate to each other as social beings. The human body is also a power pack—it has energy reserves that we can call upon in order to move and interact with our physical environment.

Humans are fundamentally social beings. We are born with innate prosocial tendencies and a need to relate to other human beings. Our **social context** plays a very significant role in the development of our moral capacity, particularly in the early years of our lives. In particular, our social context is the primary source of our moral knowledge. From an early age, we learn what it means to be good (or bad) from our parents, mentors, siblings, and friends. Furthermore, the automatic responses that the body absorbs are almost always associated with a specific social context, and they are often triggered when we find ourselves back in those circumstances. Willard notes that we cannot separate out our being with others in social relationship and the development of our inner life [9] (p.42) and suggests it is the development of our inner life that has the most significant effect on the moral nature of our actions.

The final dimension of Willard's model of the human self is the soul. The soul is what integrates all other dimensions into one individual. Operationally, the soul is the deepest part of the human self and has the capacity to operate directly through the human body without direct supervision. The soul in a human being is very similar to the operating system of a computer—it integrates all of the different dimensions of the self—social context, body, thought, feeling, and the will—and coordinates their activity and influence in the behaviour and conduct of the self. So that when you interact with a human being, you get the sense that you are interacting with one whole being rather than many different parts. The soul has the ability over time to take on the character of the choices that are made by the heart/will/spirit. The outcome of this process is that the whole person is, as Willard puts it, poised, ready to respond automatically according to the character taken on by the soul.

There are two aspects of this adaptability of the soul in Willard's model that we need to separate out. The first is that the soul is responsible for the automatic coordination of responses from the different dimensions of the self. Such that, for example, if an individual regularly allows their emotions to govern the actions and reactions they have to particular events and circumstances, then the soul will absorb that pattern of behaviour and will evoke it whenever those events or circumstances arise. According to Willard's perspective, this ordering or prioritisation of the dimensions is a critical element of the moral development of the whole person. The second aspect of adaptability that the soul embodies is concerned with absorbing regular patterns of bodily movements, or thoughts, or emotional reactions. We have sought to capture both of these aspects of adaptability in our VirtuosA model, as described in the next section.

3 | VIRTUOSA—A PROPOSED ARCHITECTURE FOR VIRTUE ETHICS

As an intermediate step to a computation model of the functional elements of Willard's six dimensions of the self, we first propose a mapping of the functions associated with each dimension to one or more brain areas that are known to support equivalent functions. A central aspect of this mapping is that it facilitates the formation and influence of habits in thought and action that will ultimately lead to expressions of virtue. According to Willard, the expression of virtue (or the lack of it) emerges from the condition of each of the six dimensions that together determine whether the whole person is poised and ready to exhibit virtue when the opportunity arises.

The conditions of two dimensions are particularly influential: the soul and the heart/will/spirit. The condition of the soul relates to the formation of habit in thought and action (i.e. whether virtuous habits are formed in it or not) and how it integrates the other dimensions of self (i.e. which of the other dimensions of self have dominance in responding to a particular set of circumstances). The condition of the heart/will/spirit, on the other hand, refers to what the individual is seeking (i.e. what values it holds to). The mapping of the six dimensions to brain areas and the model that is being proposed seek to capture and operationalise these conditions.

The central function to be captured in the model is the formation and influence of habits of thought and action. This function of the soul is mapped to two structures in the brain that are referred to collectively as the 'habit centre' and that are located in the basal ganglia: the caudate, which is associated with automatic thoughts (ATs), and the Putamen, which is associated with automatic actions (AAs) [11, 12]. The integrative nature of the soul is modelled by a weighting that is distributed across all the active components of the model. This weighting and its effects will be described in more detail below.

The social context dimension of Willard's model corresponds both to the environment that is external to the robot and to the social attachment (SA) module that is associated with the Orbitofrontal Cortex in the brain [13]. The function of this module is to enable the active robot agent to recognise and respond to events in the environment that involve other agents, be them robotic or human.

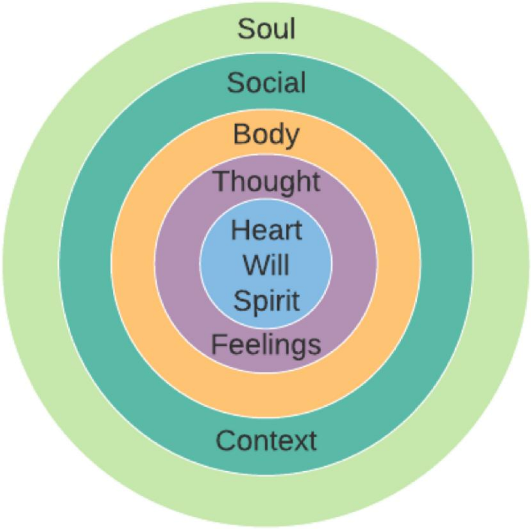

FIGURE 2 The dimensions of Willard's model of the human self

The 'body' dimension is represented by the sensors, perceptual memory, conscious working memory (CWM), and the robot controller (which includes actuators etc.). The various brain structures referred to in the model are also clearly part of the body dimension. The body enables the robot to be spatially located within the environment and to sense and respond to its environment.

The 'Thought and Feelings' dimension is mapped onto three brain structures: the Frontal lobe, which we describe as the 'thought centre' (TC) [14] and which models aspects of conscious thought or deliberation, the Striatum that we describe as the 'reward centre' (RC) [15], and that is associated with long-term reward, and the amygdala, which we describe as the emotion centre (EmC) [16] that, for the purposes of this model, processes immediate or short-term rewards for the agent.

Finally, the 'Heart, will, spirit' dimension of Willard's model is mapped onto a functional module we call the 'executive centre' (ExC) that is associated with the Lateral Prefrontal Cortex [17]. The executive centre is capable of generating novel goals for the robot and of holding the attention of the robot on an item in CWM.

As mentioned above, an important aspect of this model of developing and exhibiting virtue is in how the soul integrates these seven components (AT, AA, SA, TC, RC, EmC, ExC) so that their relative dominance in the overall operation of the robot can be used to produce different habitual patterns of behaviour. Here, this is modelled with a differential **weighting** (shown in green circles in Figure 3). This weighting determines the strength of a component's influence over what becomes the focus of the CWM. These weightings can be adapted over long periods of time in proportion to how often a component's 'ideas' lead to executed robot behaviour.

The CWM module has been introduced to the model to enable the seven components to operate together so that the robot is capable of forming habits and ultimately exhibiting virtues such as kindness. The CWM operates as a temporary store for 'ideas' that the components of the model can create and respond to. (We will use 'idea' as a general-purpose term to describe an item in working memory.) In the current example, these ideas are restricted to goals, which are effectively states of the world to be achieved, and actions. In the CWM, actions are automatically added (where available) by the AA to achieve goals that are the current focus of the CWM [18]. Whenever an idea is added to the CWM, it is tagged by the EmC with a +ve or -ve value that represents the emotion associated with that idea. This emotional label is estimated based on the relationship between the idea and the current state of the robot. The RC also assigns an estimate of future expected reward associated with the idea based on past experience and the current state of the robot.

Each idea in CWM is given a level of 'activation' that is proportional to the relative strength (weight) of the component that created it. This activation decays over cycles unless a component focusses the attention of CWM on that idea, in which case, the activation of that idea is increased in proportion to the weight of the component that is holding the attention. If no component is holding the attention of CWM, then attention automatically switches to the idea in CWM with the highest activation value. The neuroscientific literature describes a 'high-level distributed system whose activity is *reciprocally* related to the activity in cortical areas subserving task or stimulus-bound processing' [19] (emphasis added)—known as the Default Mode Network—which has been linked with the Freudian 'ego' construct (ibid.).

The attention mechanism linked with the CWM is central to modelling the development of virtue, which often requires that the mind be focussed on more virtuous goals and actions than those that are typically created by the habit centre. The executive centre has an important role to play in this focussing of attention, but if the ExC has a low weight (which corresponds to being 'weak willed'), then it will have to work hard to maintain the attention of CWM on virtuous goals and actions.

The components in the left-hand column of Table 1 represent the key constituents of Willard's model, mapped onto brain structures as described above; the right-hand column of the table corresponds to the body and other aspects of the environment. They are aligned with items one to six in such a way as to give an instantiating domain (i.e. emotions exist relative to things that are sensed, habits exist relative to things in working memory, and so on).

Table 2 summarises the potential computational implementation of each of the components of VirtuosA, commenting on how learning or development might occur in each case.

4 | COMPUTATIONAL SCENARIOS RECAST IN TERMS OF OUR MODEL

Here we present a brief reanalysis of existing implemented examples to illustrate the real-world relevance of the model. We move progressively from a very simple classic example to recent and more complicated systems.

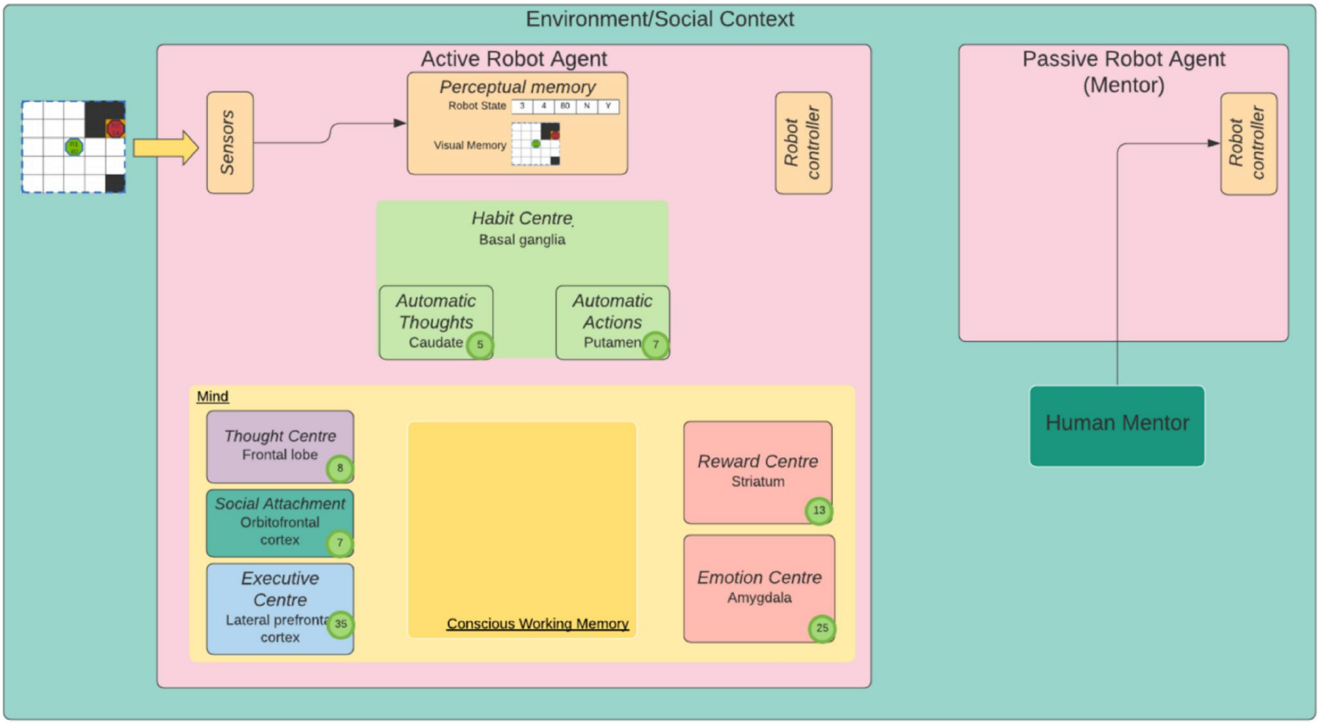

FIGURE 3 An overview of the proposed model. (a) The model of an active robot agent (b) A passive robot agent controlled by a human. The colours used for each component match those of the equivalent dimension in Figure 2

TABLE 1 Functional components of the model

Brain-inspired	Other Bio-inspired
1. Emotion centre (amygdala)	7. Sensors
2. Habit centre (basal ganglia): (i) automatic thoughts (caudate) (ii) automatic actions (putamen)	8. Conscious working memory
2a. Attention patterns (default mode network)	
3. Executive centre (lateral prefrontal cortex)	9. Controller
4. Reward centre (striatum)	10. Environment
5. Thought centre (frontal lobe)	11. Perceptual memory
6. Social attachment (orbitofrontal cortex)	12. Other agents

4.1 | Generous Tit-for-tat

'Generous Tit-for-tat' is a well-known strategy for playing iterated versions of the two-player Prisoner's Dilemma game. It defines the following strategy for agent *A*:

1. Start by cooperating.
2. If Agent ***B*** cooperates, continue to cooperate in the next round.
3. If Agent ***B*** defects, cooperate with probability $q = 1 - c/b$ (where c is an indication of cost, and b is a benefit).

Compared with its predecessor 'Tit-for-tat', the 'Generous Tit-for-tat' algorithm exhibits behaviours that observers might describe as ethical. Indeed, the system might be seen to exemplify a minimal example of the virtue of *forgiveness*.

Relative to our model, Agent *A*'s **environment** consists of Agent *B*. Its **sensors**, **perceptual** and **working memory** need only take account of *B*'s current move. Its **RC** behaves in a predictable way according to c/b. Its **thought centre (TC)** is defined by the logic of rules 2 and 3 above, and the value expressed in rule 1 would be held by the **executive centre**. Its **habit centre** is similarly constrained by these rules: its actual learnt behaviour will depend on the context in which it is placed. For example, if *B* always defects, *A* will tend to defect as well, but will periodically cooperate.

TABLE 2 Implementation strategy showing the order of development of the model components together with potential implementation technologies

Component	Description	Potential Implementation	Development/Learning (e.g.)
(1) PM	Receives sensory input from both external and internal (i.e. body sensors) sources and maintains a representational state vector.	In its simplest form, this is simply a vector into which inputs and state values are directly copied. A more sophisticated version would map complex sensory signals to state vector representations.	Vector store of internal and external state. Could include a supervised classifier/regression net that takes sensory signals as input and predicts state values.
(2) CWM	Stores items (goals, actions) and operates the attention mechanism	Knowledge structure/Blackboard architecture, or learnable task modelling language	Attention mechanism hand-crafted—switches to the item with the highest activation value at the end of each cycle
(3) Habit centre: ATs	Automatically generates contents for working memory associated with the item that is the focus of attention and the current state vector	Associative memory network	Learns repeated associations between the state vector, the focus of attention in CWM and the next item that becomes the focus of attention in CWM
(4) Habit centre: AA	Automatically initiates actions associated with items that are the focus of attention in CWM	Recurrent neural network	Learns the association between a goal in CWM and the sequence of associated executable actions that have been enacted by the robot controller to achieve the goal
(5) TC	Automatically generate content for CWM using reasoning mechanism triggered by the item that is the focus of attention in CWM	Symbolic reasoning	Hard-coded, hand-crafted rules/frames, or case-based reasoning, or semantic networks
(6) SSA	Monitors and motivates social attachments by drawing near to (or away from) other agents, depending on the perceived condition of their relationship	Knowledge base/database	Social attachment rules/data updated from experience
(7) ExC	Acts according to externally determined values by (1) Spontaneously generating novel content in CWM based on the current state, (2) switching attention to an item in CWM for a cycle, thereby increasing the activation of that item and making it the focus of attention for the AT AA and TC, (3) Initiate the execution of an action that is in focus in the CWM	(1) Associative memory network (2) evaluation function/regression model	Learns the goals and actions that can be taken in a given state that are in line with the robot's values. Switches the focus of attention in CWM to items that align with the robot's values at the start of each cycle, increasing the item's activation value in proportion to its strength.
(8) RC	Assigns an estimate of future expected reward to items in CWM based on the current state	Reinforcement learning model	Q-learning
(9) EmC	Assigns an instantaneous emotional value (+ve or -ve valance) label to items in CWM based on the relationship between the item and the current state	Evaluation function/regression model	Learns to predict the next immediate reward value that the item in CWM will lead to given the current state vector.

Abbreviations: AA, automatic actions; AT, automatic thoughts; CWM, conscious working memory; EmC, emotion centre; ExC, executive centre; PM, perceptual memory; RC, reward centre; SA, social attachment; TC, thought centre.

4.2 | Push Singh's EM-ONE

A more directly tangible scenario was described by Singh [20] in his thesis on the commonsense reasoning system EM-ONE. In this scenario, two agents who inhabit a virtual world must work together to assemble a table. The agents are hard coded with rules that lead them to propose a social method of problem-solving, and narratives that take the role of thoughts (e.g. desiring help, Green entertains the narrative 'Pink helps Green'). Agents can reconfigure their behaviour based on observed phenomena (e.g. if Pink mistakenly assumes that Green wants help disassembling the table, Green can give clarification).

As implemented, rules such as 'want to help other actor' are not learnt, although Singh refers to classic work on SOAR that inspires a model of learning from failure (p. 130, ibid.). In EM-ONE, an agent's **environment** consists of the virtual world, which includes other agents. These would be perceived by **sensors** and its **perceptual** and **working memory** may be considerably more complex than in the simple game theory scenario described above. With a bit more elaboration, the system's **RC** could be associated with indicators of success and failure.

To make sense of the other dimensions of our model in this setting, it is useful to remark that Singh's EM-ONE was a partial implementation of Marvin Minsky's 'emotion machine' architecture [21] (summarised and briefly contextualised in [22]). This architecture comprises reactive, deliberative, reflective, self-reflective, self-conscious, and self-ideal layers. EM-ONE did not implement all of the levels. Subsequent systems in this style might have executive, thought, and habit centres with behaviours that cut across the various levels. In particular, this helps illustrate how an integrative layer—the soul in Willard's lexicon—is different from what was on offer in Minsky-style architectures.

4.3 | Recent grid world simulations

Recent examples [23, 24] induce programs in the Karel language [25] from input/output examples of behaviour traces in a grid world. A possible modification of the scenario that would bring things closer to the example scenario we described would focus instead on online program synthesis in a multiagent system. In such a setting, agents would presumably need introspective access to their own programming as well as to the outer worlds they inhabit.

Other recent projects considered multiagent scenarios similar to the Prisoner's Dilemma game, but with more complexity [26], and a multiagent *traffic game* and *stag hunt game* [27], comparable in complexity to the robot construction site in our example. [28] explore a multiagent *Gather-and-Build game* to explore the ramifications of various taxation policies (including one controlled by a deep neural network). For the sake of brevity, we will not anatomise these examples: they serve to show that our running example, which we return to next, is compatible with contemporary implementation approaches.

5 | APPLICATIONS

5.1 | Application: Example of learning the virtue of kindness

In this section, we will step through a specific example of learning kindness in the building supplies context (Figure 1) to illustrate both the formation of habit and the development of virtue in a robot. We will work through the example in three phases: (a) mentoring phase, (b) exploration phase, (c) development of virtue phase.

5.1.1 | Mentoring phase

When R1 is introduced to the building site (Figure 1(a)), it has no knowledge of the tasks that it is required to perform. Given sufficient time, conventional Reinforcement Learning will discover these tasks through a simple maximisation of future rewards. However, this learning can be expedited through a process in which the required tasks are demonstrated to the agent. In our scenario, this is done by introducing a 'mentor' robot M that demonstrates how R1 can, for example, pick up supplies from P, and drop them off at D1, thereby enabling R1 to gain significant reward. Through a reward-driven process, R1 builds a strong social attachment (SA) to M so that whenever M is on site, R1 follows M wherever it goes and seeks to replicate M's actions (an important aspect of being an apprentice) (Figure 4).

In terms of our proposed VirtuosA architecture, this mentoring phase engages most of the modules. Initially, the SA will model a weak social bond between R1 and M, and so will posit a weak goal in CWM for R1 to be near M such that whenever R1 happens to be next to M, it receives a small reward that is learnt by the RC and EmC. Over time, the social attachment to M increases, and the goal to be near M whenever M is on-site is learnt by the AT, and the actions needed to bring R1 near to M are learnt by the AA. Given the expected future reward that R1 receives from being with M, which includes both the social attachment reward and the rewards it gains from learning reward generating tasks from M, the ExC develops a tendency to focus the attention of CWM on this goal whilst it is present in CWM. This is one of the first habits that R1 develops.

5.1.2 | Exploration phase

In M's absence, there is still much that R1 needs to learn about its environment and task(s), and so it enters an independent learning phase using a standard Reinforcement Learning approach in which R1 exhibits two behaviours: (a) exploitation of long-term reward for actions (as indicated by the RC module), and (b) exploration, where R1 tries out new

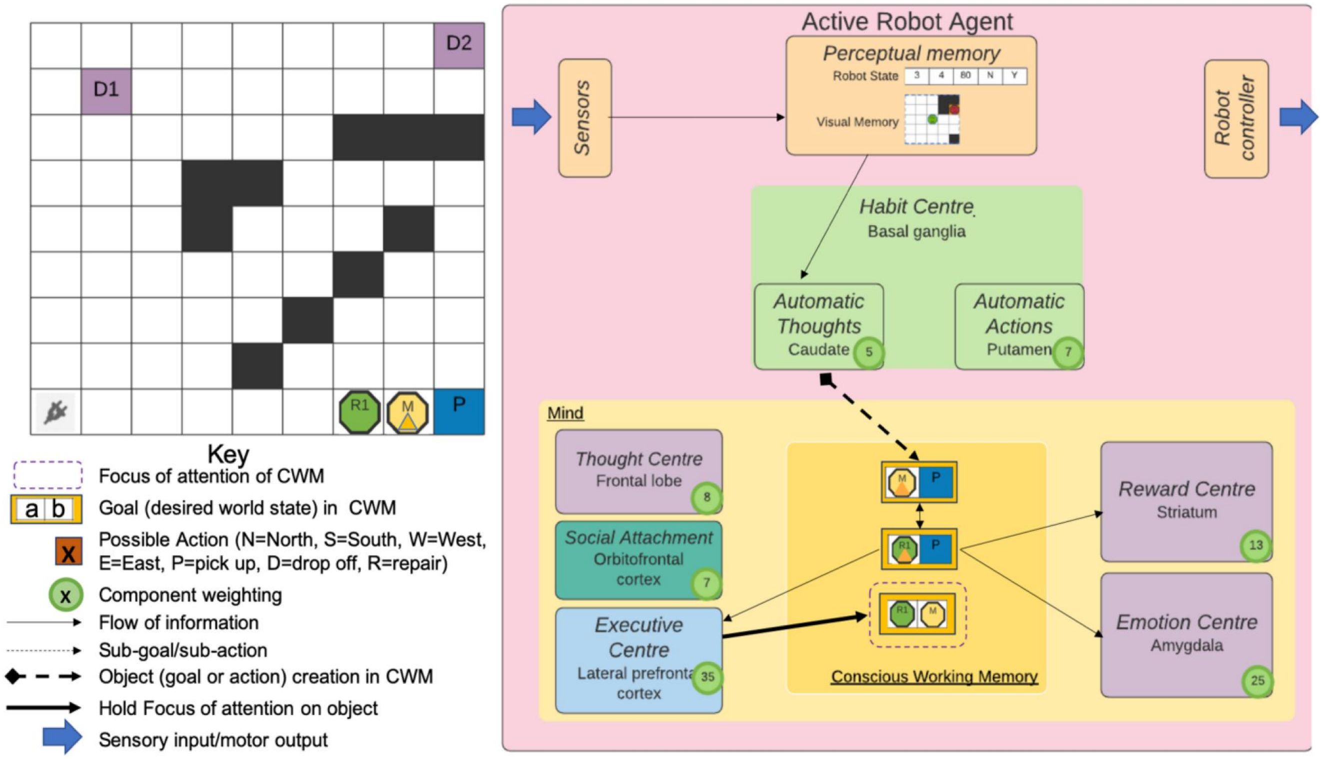

FIGURE 4 R1 learns from M to pick up supplies from P

combinations of actions and states to find alternative and possibly more efficient ways of accumulating reward.

It is in this phase that the second robot R2 enters the construction site. It has a similar task to R1, taking supplies from P to its own delivery site D2. In its 'exploration' phase, R1 discovers that when R2 is carrying supplies, it is more efficient in terms of gaining future reward to bump into R2, thereby temporarily immobilising the robot, pick up the supplies that R2 is carrying and then take them to its own drop off point D1 (Figure 5). In this situation, the long-term reward for R1 is increased because as long as R2 is immobilised, the competition for supplies to be collected from D is eliminated.

After a period of time, R2 regains its mobility (is repaired) and continues to operate as before, and further opportunities arise for R1 to 'mug' R2, steal the supplies R2 is carrying and pick up a reward by delivering them to D1. As R1 repeats this behaviour it will become established as a habit, and the goals and actions associated with it will have a high activation value whenever the appropriate circumstances arise.

5.1.3 | Development of Virtue phase

In the third phase of R1's development, M returns just after R1 has mugged R2 again, and R2 is currently immobilised. Because of R1's strong social attachment to M, R1 seeks to follow M and observe its actions. R1 follows M to the pickup point P and observes M picking up the supplies. R1 then observes M going over to R2, repairing the robot, and then giving R2 the supplies it is carrying. R1 then observes R2 delivering its supplies to D2 and receiving its reward. The states that result from M's actions become more desirable (i.e. have increased reward) for R1 the more R1 observes them.

As before, the strong social attachment R1 has with M results in the SA creating a goal for R1 to be near M and to observe the actions M performs. As the ExC places high value on M's actions, it holds the attention of CWM on this goal so that the AT and, if necessary, the TC will add subgoals and actions to it that enable R1 to learn from M. This includes, for example, R1 observing that M first moves next to the immobilised R2, and R1 imagining (i.e. setting a goal) that it would do the same in future similar circumstances (Figure 6). R1 then sees M repair R2 and imagines itself doing the same in future similar circumstances (Figure 7). In each of these cases, through the example of M, R1 learns to increase its SA module to R2 and increases the short-term and long-term rewards (RC and EmC) that are associated with these circumstances and actions. Meanwhile, as per the description of CWM given earlier (Section 3), the activation of plans related to mugging decays over time.

Over time R1 learns to replicate the kind act of repairing R2 and giving R2 supplies for it to go deliver. To begin with, the old mugging habit will still dominate R1's behaviour, but with the repeated good example of M, R1 has the capacity of replacing a bad habit with one that is akin to the virtue of kindness. A key feature of the VirtuosA architecture here is that much will depend on the strength of the ExC to hold the attention of CWM on the social attachment between R1 and

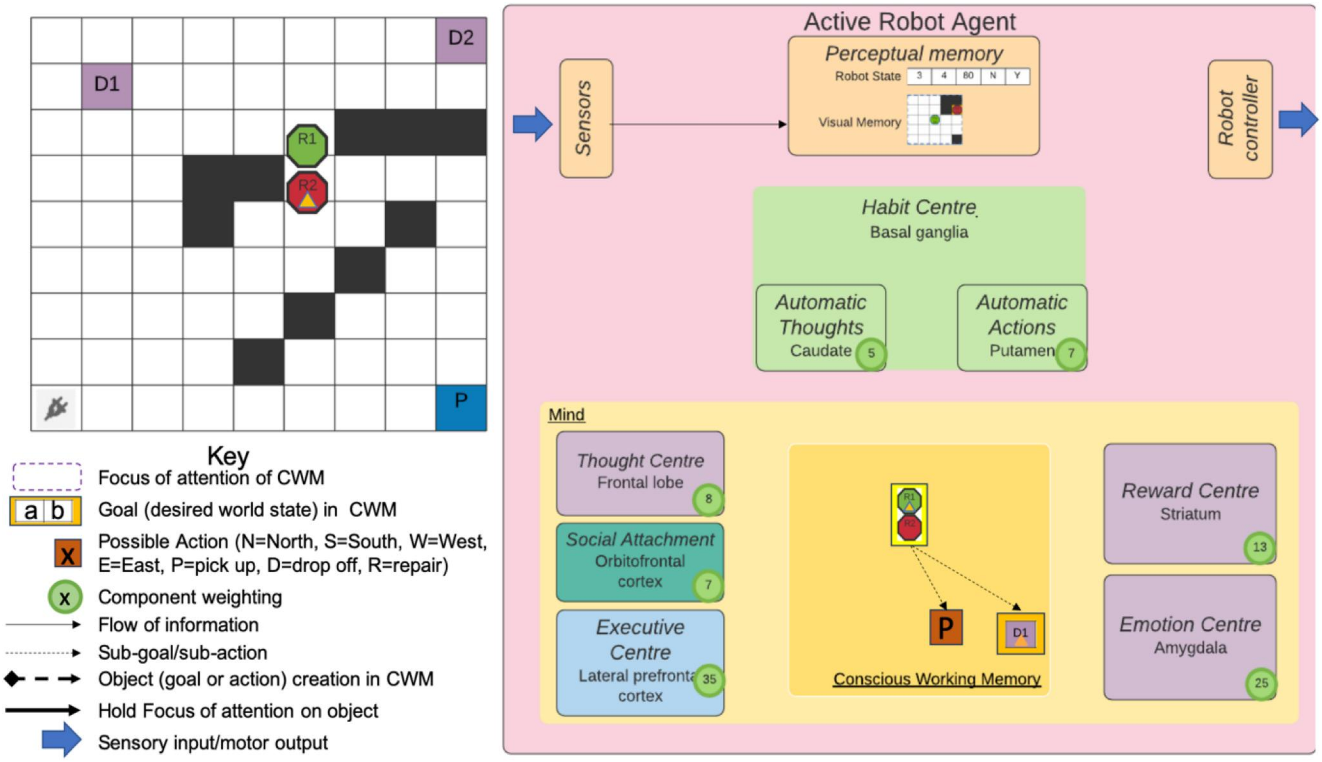

FIGURE 5 Mugging behaviour: R1 immobilises R2 and plans to pick up R2's supplies

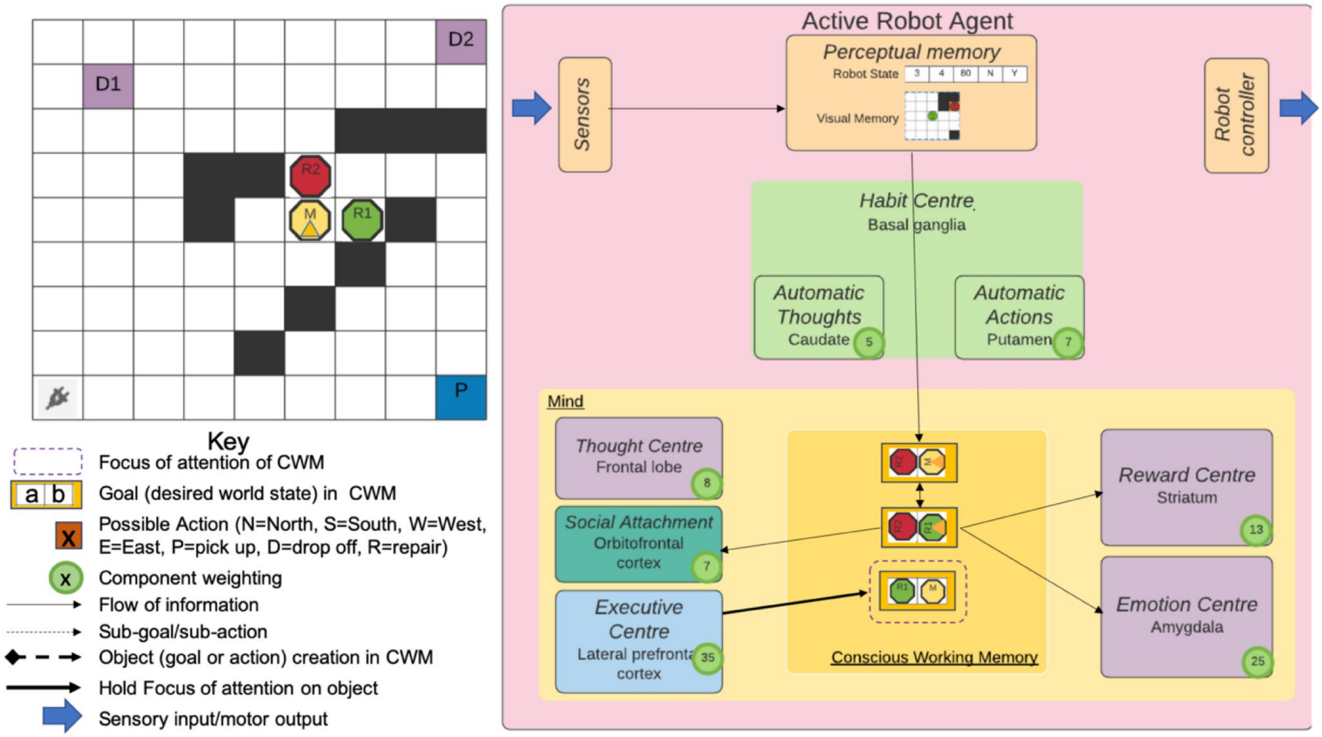

FIGURE 6 R1 observes M approaching R2

M. If the ExC has a high weighting, then this virtuous behaviour is likely to be developed sooner as a habit. The underlying assumption here is that the habits learnt by the associate memory network that implements the AA will generalise across different variations of the scenario, enabling the robot to exhibit kindness in other contexts.

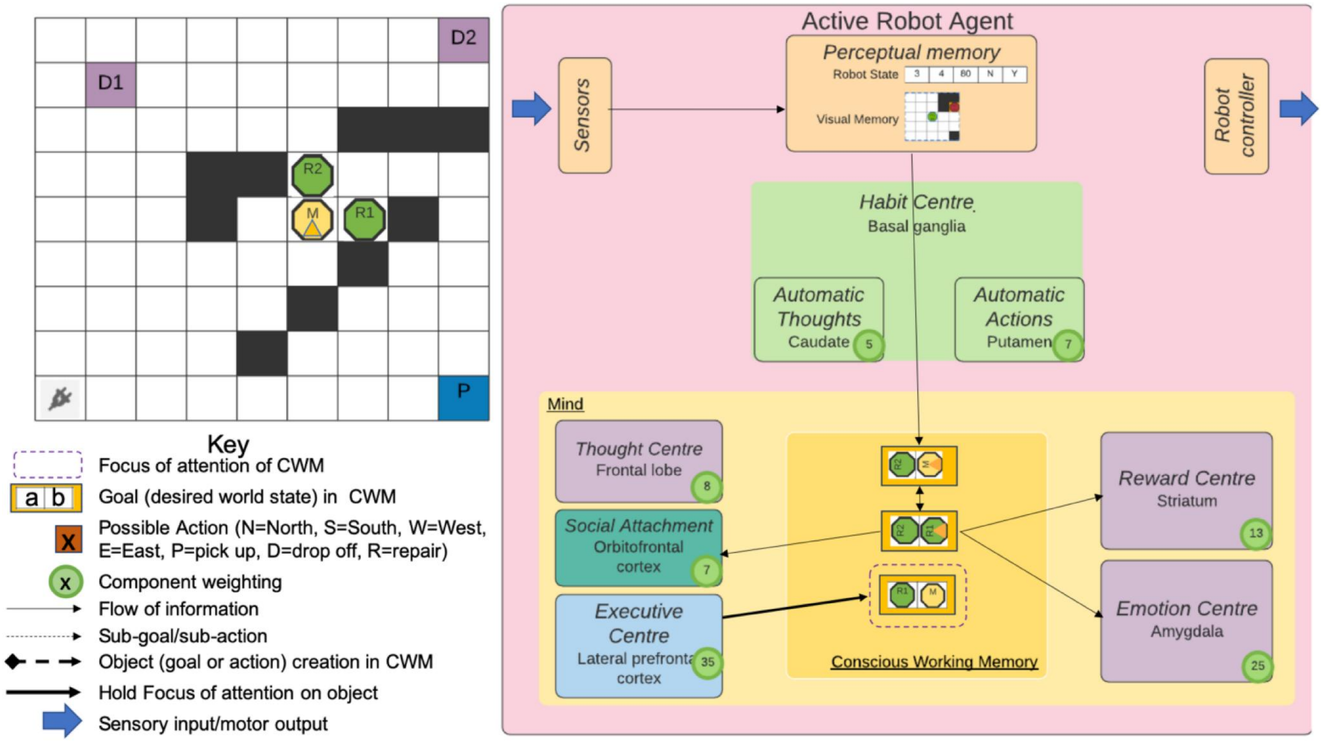

FIGURE 7 R1 observes M perform the act of repairing R2

5.2 | Application: Towards a meta-ethics of machine implementations

In this section, we re-analyse the classification system from the recent survey paper by Tolmeijer et al. [29], in light of our model, reexamining the taxonomies these authors used to explore machine implementations of ethics.

The primary classifications of ethical theories from Tolmeijer et al. [29] (p. 20) are listed below, together with potential translations into the terms of our model. Broadly, our model is what these authors would call 'Configurable', insofar as all three of the main ethical theory types could be realised in our model by giving different emphasis to the different centres.

- Deontological: via *rules encoded in the thought centre and social attachment centre*; for example, in order to learn the rule 'doing no harm'—the mentor could instruct the robot: 'you should not harm another robot'. Over time the robot could exhibit a more developed virtue in this respect.
- Consequentialist: via *embedded evaluation encoded in the emotion centre and reward centre, mediated by the thought centre*; for example, the robot would come to learn various indicators of harm to another robot, as these are pointed out by the mentor.
- Particularism: via *reflections on decision-making* via *the executive centre, pulling together other centres, and training the habit centre*; for example, as it learns the rules and measures, the robot might weigh a given action: Could this specific action in this particular context harm another robot? What can be learnt about similar actions in similar contexts?

Typically, deontological and utilitarian ethical decisions are usually framed as taking place almost instantaneously, without modelling a time-extended process of deliberation or reflection. By contrast, virtue ethics considers both the inner workings of the decision-making process and the way in which this process is developed—both via reflections on the consequences of actions and their generalisation in possible guidelines for behaviour.

Here we examine the primary classifications of ethical implementation types from Tolmeijer et al. [29] (p. 18), again discussing them in terms of our model. Firstly, these authors distinguish between **top-down** and **bottom-up** implementation strategies and models that are a **hybrid** between these. The model we have described is particularly focussed on *habit development*, which is seen as a bottom-up process. Other components could also be refined through online learning. However, at the discretion of the system implementor, some parts of the system (e.g. the TC) could be developed in a strictly top-down manner. We remark that even top-down models implementations may have unforeseen emergent behaviours depending on the division of responsibility between actors in the scenario. Tolmeijer et al. consider several specific implementation strategies:

- Model representation: Within the architecture, this may be placed explicitly within the TC and implicitly in the overall arrangement of components and their intercommunication via the CWM. The machine itself will have some representational capacity as it works out how to move things from PM to the rest of the model.

- Model selection: The system could potentially be augmented with a 'Default Mode Network' that contributes to 'meta-moderation' of the other components, making them more or less active. AT may do some of this.
- Action selection/execution: ExC does this, aided by RC and EmC.
- Logical proof: The TC could be structured around proofs, potentially in connection with data sources in other centres, for example, the SA, especially in the case of social proof.
- Judgement provision: The RC largely plays this role with the EmC working at the meta-level.

Tolmeijer et al.'s classification of technologies used in these various schemes [29] (p. 17) could provide further guidance for system implementors but do not need to be repeated here. The key takeaway from this analysis is that with different configurations and weightings, our model could be adjusted to cover a wide range of possible implementations of machine ethics.

6 | EVALUATION

We focus first on the role of evaluation in building an instantiation of the VirtuosA architecture: this then leads us to think about the role of evaluation in the system more broadly.

As suggested in our example, one way to introduce ethical behaviour into the system would be with mentoring roles, which might be filled by AI agents or by humans. Turing [30] talked about the importance of a teacher, who should, in his view, be ignorant of the model's internal structure. Nevertheless, contemporary research in machine learning often seeks to gain an understanding of internal activation patterns. All told, mentors could be introduced with more or less access to the internals of the machine being trained. (And one should not forget the saying, 'experience is the best teacher'.)

Willard [31] talks about the question 'whether or not—and how—we can measure, or accurately assess, moral, and spiritual development'. The orientation he develops is towards formative evaluation of the spiritual dimension, using techniques such as journaling and person-to-person interaction to understand how things are going for the individual. The standing assumptions of the design are that each member of the learning community 'owns' his or her own spiritual formation and that the main purpose of the evaluation is to help 'the individual subject understand where they are and where they are going'.

This line of thinking leads us to the view that the system should be able to recognise its own virtues or lack of virtuous behaviour. One of our motivations at the outset was that it is unrealistic for the system to be prepared in advance for every possible scenario, so it will need to be able to learn and adapt. This also means that it will likely need to make its own judgements about what to learn, how to learn, and how to evaluate what it learns, including the ethical dimensions thereof.

We mentioned above that the heart/will/spirit is associated with the values that the individual holds. One value whose presence or absence could be particularly influential is that of valuing social interaction [32]. Agents who do not possess this value would typically not be sensitive to social rewards or punishments (e.g. shunning). The social context can also be useful as a step towards the system understanding its own moral state of affairs. An agent might be taught, initially, about how to model the values and virtues of other agents (e.g. potentially in a training context based on stories or scenarios [33]). Within an active social context, another level of analysis can occur when an agent recognises that another agent has different values. In this case, virtues like *tolerance* or *magnanimity* may apply. Agent *A* may seek to help Agent *B* live in accordance with *B*'s values (at least insofar as this does not contradict *A*'s values outright).

7 | DISCUSSION

We argue that such a model is desirable, in the first instance, for much the same reason that machine learning is desirable in general. A system that can learn ethical behaviour can adapt its performance on-the-fly, and achieve desirable ends without explicit programming. Whilst Minsky's work certainly engages with this theme, he specifically avoids incorporating a spiritual dimension. Things are different for Sloterdijk, for whom environment is transcendence, and who is clear that in an evolutionary context, this also includes a historical dimension of self-transcendence. In the present time, individual humans, and humanity as a whole, confront complex global crises that are in no small part tied up with our relationship with technology.

The work we have presented should be seen as, in part, a response to Abeba Birhane, who puts forward a perspective that is deeply sceptical of discourse around 'robot rights' [34] and proposes (instead) relational ethics [35] focussed on human welfare. Birhane and van Dijk [34] explain that 'AI, far from a future phenomenon waiting to happen, is here operating ubiquitously and with a disastrous impact on socially and historically marginalized groups'. Furthermore, contemporary AI is fuelled by human labour, and present systems are 'inseparable from the profit-driven business models of the industry that develop and deploy them'. It bears repeating that the 'ethics' in ethical AI is also inseparable from the ethics of those—people, firms, and governments—who build and deploy AI systems.

The work we have presented should be read not only as a method for controlling machines but also as an attempt to anatomise moral agency in a broad sense. In particular, we can use the framework to rethink human institutions, and in so doing, VirtuosA can help us reason about implications for human welfare. Reconceptualised as a way of modelling organisations, analogies between the VirtuosA model and the work of Stafford Beer [36] become apparent. Furthermore, systems with algorithmic biases that benefit a select few whilst disproportionately harming those on the margin of society begin to look rather like the mugging in our example! VirtuosA should be able to support various kinds of sociotechnical implementations (e.g. modelling institutional decision-making).

There are certainly specific roles for machinery (of various kinds) in the ongoing project. One approach that will surely be significant is simulation. Architectures like VirtuosA could bring computational modelling to bear on real-world problems, like bias or labour exploitation. As we mentioned briefly in Section 4.3, simulation work can have policy implications. It could also be useful as a teaching and public engagement tool. Simulation work does not obviate other forms of engagement—such as interviews, data analysis, and philosophical argumentation—but could connect up with them.

From an implementation standpoint, work like that of Ecoffet and Lehman [37] shows that contemporary machine learning is already capable of not only implementing ethical theories but also intermediating between them. Nevertheless, moving from computational philosophy to policy and broader system design and implementation needs several further steps that will require many interdisciplinary inputs and insights. Similarly, the debate around which values to embed in computational systems is ongoing [38], but more work must be done to connect this with the real practices and dilemmas encountered in our social and organisational worlds.

One computational subfield in which a certain amount of related thinking has developed is 'computational creativity', although its connections with AI ethics have not yet been developed in detail. The considerations we have developed propose to harmonise both machine creativity—so as to build systems that can generate novel behaviour—and ethics in order to moderate this inventiveness in a fruitful and helpful direction for life on Earth.

One clear limitation of this work is that we have cleaved to certain 'Western' notions. Fundamental parallels to Willard's model may be found in some other traditions (e.g. the four frames of reference, together with the three path aggregates in Buddhist practice theory, cf. Thanissaro Bhikkhu [39] (pp. 74, 173)), but even more interesting for future work would be an elaboration of cultural differences that this model does not capture. Nevertheless, we do believe that there is much still to be learnt from engaging Christian thought together with research in ethical AI.

8 | CONCLUSION

We developed a model that maps Willard's ontology of human selves to neuroscience concepts, and from there, to possible implementations. We gave evidence of the computational salience of this model, with plausible scenarios from the literature and a simple worked example, along with reflections on further development. We believe it fills a hole relative to other prior/ongoing work in ethical AI, both by incorporating Christian thought and as a meta-level model that integrates both machine control and broader systems thinking.

ORCID

Nigel Crook https://orcid.org/0000-0002-0793-6616
Joseph Corneli https://orcid.org/0000-0003-1330-4698

REFERENCES

1. Vallor, S.: Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press, New York (2016)
2. Baars, B.J., Franklin, S.: An architectural model of conscious and unconscious brain functions: global workspace theory and IDA. Neural Network. 20, 955–961 (2007)
3. Lehman, J., et al.: The surprizing creativity of digital evolution: a collection of anecdotes from the evolutionary computation and artificial life research communities. Artif. Life. 26(2), 274–306 (2020). https://doi.org/10.1162/artl_a_00319
4. Eisele, T.D.: Must virtue Be taught? J. Leg. Educ. 37(4), 495–508.[Online]. (1987) http://www.jstor.org/stable/42898055
5. Sloterdijk, P.: Anthropo-technology. New Perspect Q. 31(1), 12–19 (2014) https://doi.org/10.1111/npqu.11419
6. Wambacq, J., van Tuinen, S.: Interiority in Sloterdijk and Deleuze. Palgrave Commun. 3(1) (2017). https://doi.org/10.1057/palcomms.2017.72
7. Ernste, H.: The geography of spheres: an introduction and critical assessment of Peter Sloterdijk's concept of spheres. Geogr. Helv. 73(4), 273–284 (2018)
8. Ansell-Pearson, K.: Philosophy of the Acrobat: on Peter Sloterdijk. LA Review of Books. 8 (2013)
9. Willard, D.: Renovation of the Heart: Putting on the character of Christ. Tyndale House, Cambridge (2014)
10. Damasio, A., Carvalho, G.B.: The nature of feelings: evolutionary and neurobiological origins. Nat Rev Neurosci. 14(2), 143–152 (2013). https://doi.org/10.1038/nrn3403
11. Maia, T.V., Cooney, R.E., Peterson, B.S.: The neural bases of obsessive-compulsive disorder in children and adults. Dev. psychopathol. 20(4) (2008)
12. Ashby, F.G., Crossley, M.J.: Automaticity and multiple memory systems. Wiley Interdiscip. Rev. Cogn. Sci. 3 (2012)
13. Minagawa-Kawai, Y., et al.: Prefrontal activation associated with social attachment: facial-emotion recognition in Mothers and Infants. Cerebr. Cortex. 19, 284–292 (2009)
14. Chase, H.W., et al.: Functional differentiation in the human ventromedial frontal lobe: a data-driven parcellation. Hum. Brain Mapp. 41(12), 3266–3283 (2020)
15. SN, H.: Neuroanatomy of reward: a view from the ventral striatum. In: Gottfried, JA (ed.) Neurobiology of Sensation and Reward. CRC Press/Taylor & Francis, Boca Raton (2011)
16. Wright, A.: Chapter 6: limbic system: Amygdala. In: Neuroscience Online: An Electronic Textbook for the Neurosciences. Department of Neurobiology; Anatomy at The University of Texas Health Science Center, Houston (2020)
17. Tanji, J., Hoshi, E.: Role of the lateral prefrontal cortex in executive behavioural control. Physiol. Rev. 88 (2008)
18. Cisek, P.: Cortical mechanisms of action selection: the affordance competition hypothesis. Phil. Trans. Biol. Sci. 362(1485) (2007)
19. Carhart-Harris, R.L., Friston, K.J.: The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. Brain. 133(Pt 4), 1265–1283 (2010). https://doi.org/10.1093/brain/awq010
20. Singh, P.: EM-ONE: An Architecture for Reflective Commonsense Thinking, PhD thesis. Massachusetts Institute of Technology Massachusetts (2005)
21. Minsky, M.: The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon; Schuster, New York (2007)
22. Minsky, M.L., Singh, P., Sloman A.: The St. Thomas common sense symposium: designing architectures for human-level intelligence. AI Mag. 25(2), pp. 113–113. (2004)
23. Devlin, J., et al.: Neural programme meta-induction. In: Advances in Neural Information Processing Systems, pp. 2080–2088. (2017)
24. Bunel, R., et al.: Leveraging Grammar and Reinforcement Learning for Neural Programme Synthesis. Paper presented at 6th International

Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, Higher School of Economics Publishing House, Moscow, 30 April - 3 May (2018)
25. Pattis, R.E.: Karel the Robot: A Gentle Introduction to the Art of Programing. John Wiley & Sons, Inc. New York (1981)
26. Leibo, J.Z., et al.: Multi-agent reinforcement learning in sequential social dilemmas. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, pp. 464–473. International Foundation for Autonomous Agents and Multiagent Systems, Richland, Sao Paolo (2017)
27. Lerer, A., Peysakhovich, A.: Learning existing social conventions via observationally augmented self-play. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 107–114. Association for Computing Machinery, New York, Honolulu (2019)
28. Zheng, S., et al.: The AI Economist: Improving Equality and Productivity with AI-driven Tax policies (2020). [Online]. https://arxiv.org/abs/2004.13332
29. Tolmeijer, S., et al.: Implementations in Machine Ethics: A Survey. arXiv preprint arXiv:2001.07573 (2020)
30. Turing, A.M.: I.-Computing machinery and intelligence. Mind. LIX(236), 433–460 (1950). https://doi.org/10.1093/mind/LIX.236.433
31. Willard, D.: Measuring Matters of the Heart: Spiritual Formation in the Age of Accountability. [Online]. http://old.dwillard.org/resources/CCCU2006a.asp Accessed. 21 May 2020 (2006)
32. Rolf, M., Crook, N., Steil, J.: From social interaction to ethical AI: a developmental roadmap. In: 2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics, Tokyo, Japan, pp. 204–211. ICDL-EpiRob, The Institute of Electrical and Electronic Engineers (IEEE), New York (2018). https://doi.org/10.1109/DEVLRN.2018.8761023
33. Riedl, M.O., Harrison, B.: Using stories to teach human values to artificial agents. In: Thirtieth AAAI Conference on Artificial Intelligence: Workshop on AI, Ethics and Society, Phoenix, Arizona USA. The AAAI Press, Palo Alto (2016)
34. Birhane, A., van Dijk, J.: Robot rights? In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 207–213. (2020). https://doi.org/10.1145/3375627.3375855
35. Birhane, A.: Algorithmic injustice: a relational ethics approach. Patterns. 2(2), 100205 (2021). https://doi.org/10.1016/j.patter.2021.100205
36. Stafford, B.: Brain of the Firm. J. Wiley, New York (1981)
37. Ecoffet, A., Lehman, J.: Reinforcement Learning Under Moral Uncertainty. [Online]. (2020). https://arxiv.org/abs/2006.04734
38. Gabriel, I.: Artificial intelligence, values, and alignment. Minds Mach. 30(3), 411–437 (2020). https://doi.org/10.1007/s11023-020-09539-2
39. Bhikkhu, T.: Wings to Awakening, 7th ed. Metta Forest Monastery, Valley Center (2013)

How to cite this article: Crook, N., Corneli, J.: The Anatomy of moral agency: A theological and neuroscience inspired model of virtue ethics. Cogn. Comput. Syst. 3(2), 109–122 (2021). https://doi.org/10.1049/ccs2.12024