# The Role of Cognitive Architectures in the Modelling of Human Virtue

Fraser Watts

**Abstract:** There has been increasing interest in virtue, both from the perspective of theology, and from human sciences such as psychology. This paper focuses on the possibility of modelling human virtue computationally, in terms of a particular cognitive architecture, Philip Barnard's Interacting Cognitive Subsystems. It is unusual in being a macro cognitive architecture; it has been formulated with a computational level of precision; and its evolutionary development has been described. Though it is possible that computers could acquire virtue in a distinctive way, the focus here is on modelling human-like virtue. That needs to be grounded in empirical research on human moral functioning, and research is discussed on two main topics: whether or not virtue functions as a cross-situational, characterological trait; and the tendency of virtue to fragment into components such as values and behaviour which are only weakly connected. An important feature of Interacting Cognitive Subsystems is the distinction between two different modes of central cognition, one intuitive-embodied and the other conceptual. The interplay between these two different modes of cognition in the acquisition of virtue is discussed.

**Keywords**: behaviour; conceptual cognition; intuition; modelling; values; virtue

Fraser Watts, formerly Reader in Theology and Science in the University of Cambridge and President of the British Psychological Society, is now Executive Secretary of the International Society for Science and Religion and Visiting Professor of Psychology of Religion at the University of Lincoln. His latest book is *A Plea for Embodied Spirituality: The Role of the Body in Religion* (2022).

*Christian Perspectives on Science and Technology*, New Series, Vol. 3,
Special Issue: Artificial and Spiritual Intelligence (2024), https://doi.org/10.58913/RZUV8384

There has recently been increasing interest in virtue, both in Christian ethics and in moral philosophy. The traditional Christian approach to moral issues is the deontological one, framed in terms of moral laws and norms. In the latter part of the 19th century that was supplemented by a consequentialist approach, of which utilitarianism is the best developed example. In colloquial terminology, the alternative approaches are "follow the rules," or "be nice to people." However, there has been growing dissatisfaction with these alternatives among both philosophers and Christian theologians. As Rowan Williams puts it, neither "crisp moral precision" nor "gentle vagueness about not hurting people" seems to be an adequate approach. He asks, "I wonder what exactly is supposed to be *Christian* about either of them?"; and comments, "Christian morality is always lurching between a touching faith in the power of rules to secure your place with God, and a rather vacuous reliance on 'inner' convictions and sincerity."[1]

An alternative approach to Christian ethics, that has ancient roots but which became increasingly influential in the 20th century, is framed in terms of virtue or character.[2] Philosophically, the shift to virtue in ethics was associated with a new emphasis on human action being inherently intentional. An emphasis on virtue also connects well with theological anthropology, with its rich analysis of the human predicament recognising how difficult it is for humans to get things right, but the importance of aspiring to do so. If humans are to behave well at all, that arises from a long and sustained process, whether it is framed in terms of cultivating virtue, developing character, or educating conscience.

The theological approach will emphasise the importance of relationship with God. However, there is a complementary approach from the human sciences centred around how virtue is cultivated. This is the focus of this paper, which will propose an approach to the cognitive

[1] Rowan Williams, "Is There a Christian Sexual Ethic?" in Rowan Williams, *Open to Judgement: Sermons and Addresses* (London: Darton, Longman and Todd, 1994), 162–163.

[2] Edward Leroy Long Jr, *A Survey of Recent Christian Ethics* (New York: Oxford University Press, 1982).

*Christian Perspectives on Science and Technology*, New Series, Vol. 3,
Special Issue: Artificial and Spiritual Intelligence (2024), https://doi.org/10.58913/RZUV8384

modelling of human virtue using a particular cognitive architecture. Cognitive modelling is a very precise mode of theorising, and its rigour and precision have been a considerable asset in many areas of psychology. It will help us to understand human virtue better and to specify in the language of a cognitive architecture the psychological processes through which virtue arises.

My focus here is on *human* virtue. In principle, there might be other approaches to being virtuous besides the human one; computers might approach virtue differently from humans. AI has been very successful in developing *human-level* intelligence in computers, but it generally achieves intelligence that is human-level but not *human-like.*[3] AI often takes advantage of the strengths of computers to circumvent their lack of certain human capacities. Computers are very good at many things, like playing chess, but they generally do them differently from humans.

To simulate human virtue computationally, one needs to start with the empirical study of human moral functioning. Understanding how morality (and virtue) work in human beings requires a broadly-based human science such as psychology. In this paper I will nest what I say about virtue within a broader consideration of human moral functioning (just as virtue ethics is one particular approach to ethics). Being virtuous can be regarded as a particular approach to being moral. So, I will begin with the broader topic of the psychology of morality before coming to the more specific topic of the psychology of virtue. I do this partly because psychology has more to say about morality than about virtue.

There are three issues arising from the psychology of virtue that I want to highlight, because of their implications for the computational implementation of virtue. First, virtuous behaviour is more situationally specific, and less governed by characterological traits, than is

[3] Marius Dorobantu, "*Imago Dei* in the Age of Artificial Intelligence: Challenges and Opportunities for a Science-Engaged Theology," *Christian Perspectives on Science and Technology*, New Series, Vol. 1 (2022): 175–196, https://doi.org/10.58913/KWUU3009.

commonly assumed. Second, virtue is multifaceted, and the various components of virtue don't hang together as much as is commonly assumed. Third, I will claim that virtue is inherently relational.

Later in the paper, I will suggest an approach to cognitive modelling of virtue through a rigorously-specified cognitive architecture, focusing particularly on Philip Barnard's Interacting Cognitive Subsystems (ICS).[4] I suggest that working towards the computational implementation of virtue through theoretical work at the level of a computational-level cognitive architecture is a helpful intermediate step between empirical research on human virtue and full computational implementation of human-like virtue. If we try to jump over that intermediate stage we may, at best, end up with replication, not simulation, and with a computational approach that is not human-like.

## Is Virtue a Character Trait?

It is one of the bedrock assumptions of the virtue approach to ethics that virtue is a matter of character, and that virtuous people are consistently virtuous.[5] N. T. Wright uses the analogy with a stick of Brighton Rock and says that, with virtue, wherever you cut it you find the same words in the centre.[6] It is assumed that over a long period of time people can develop their character or, in the terminology of a previous age, "educate their consciences," in a way that will make them more virtuous. The assumption is that if people are virtuous at all they will be consistently so.

Is this assumption correct? Whether people are consistently virtuous across different situations and contexts is a complex matter to investigate empirically, and there has been very little attempt to carry it out. One classic study by Hartshorne and May looked at truthful-

[4] John D. Teasdale and Philip J. Barnard, *Affect, Cognition and Change* (Hove: Lawrence Erlbaum, 1993).

[5] Harris Wiseman, *The Myth of the Moral Brain: The Limits of Moral Enhancement* (Cambridge, MS: MIT Press, 2016).

[6] N. T. Wright, *Virtue Reborn: The Transformation of the Christian Mind* (London: SPCK, 2010).

ness in children at home and in school.[7] If children are truthful, can they be relied on to tell the truth in all situations, or does truthfulness depend on the context? The research found only very weak correlations between truthfulness at home and at school. At least in children, truthfulness does not seem to be a virtue that transcends the particular situation.

A landmark discussion of these issues was *Personality and Assessment*, by Walter Mischel,[8] which argued convincingly that personality was much more situationally specific than had generally been assumed. People seem quite wedded to the assumption that personal behaviour is trait-like, and that people are consistent in how they behave across situations. That assumption is strong, both in folk psychology and in virtue ethics. However, the possibility needs to be considered that this widespread assumption is actually wrong, and that virtuous behaviour is much more variable from one context to another than is normally recognised. This is something that Blaine Fowers has noted in his science of virtue.[9] He notes that traits are assumed to show consistency in behaviour, cognition and affect over time, but he also notes that many studies on virtue only collect data at a single point in time, so cannot assess consistency. There has recently been growing interest in what can be learned about virtue from studying variation over time in the same individuals, rather than differences between individuals.

If people are virtuous in some situations but not others, some kind of appraisal process must go on, at least unconsciously, of which people have an intuitive awareness. Such appraisals determine how people behave in a particular situation, and will need to be incorporated in a computational implementation of human-like virtue. However, we have only a limited formal, scientific understanding of how such appraisals work in humans, so we don't know how to model

[7] H. Hartshorne and M. A. May, *Studies in Deceit. Book I: General Methods and Results. Book II: Statistical Methods and Results* (London: Macmillan, 1928).

[8] Walter Mischel, *Personality and Assessment* (New York: John Wiley, 1968).

[9] Blaine J. Fowers, "Toward Programmatic Research on Virtue Assessment: Challenges and Prospects," *Theory and Research in Education* 12:3 (2014): 309–328, https://doi.org/10.1177/1477878514546064.

them in computers. Taking the example of truthfulness, it may be partly a judgement about how important it is to tell the truth in a particular situation, and partly a judgement about what people can get away with in a particular situation.

An alternative approach to the computational modelling of virtue might be to say that this situational specificity in human virtue is an unfortunate human characteristic. It might be argued that computers are capable of showing virtue consistently, even though humans normally fail to do so. It is possible, of course, that exceptionally virtuous people are more consistent in how they behave across situations than the majority of people. Moral consistency may be an aspiration which, in some exceptional people, may become a reality.

## The Multifaceted Nature of Virtue

The next issue about moral functioning to be considered is that it is multifaceted, and is often fragmented. Moral goodness and virtue are not a single thing. They have various different aspects. As with many other aspects of human functioning, it is important to distinguish thoughts, feelings, and actions, at the very least. That is not unique to virtue or morality. It is true of most high-level human functioning, including religion and spirituality. For example, to understand religion psychologically it is necessary to consider religious understanding and beliefs, religious feelings and experiences, and religious behaviours and practices.[10]

There is sometimes a tendency to simplify matters by picking on one particular facet, and to say that all that really matters is to simulate that computationally. Most often, it is *behaviour* that people seize on and say, for example, that if a computer exhibits emotional behaviour, it has emotions. I accept that, for some practical purposes, it is enough to replicate performances. However, I claim that it is inherent in the

[10] Fraser Watts, *Psychology, Religion and Spirituality: Concepts and Applications* (Cambridge: Cambridge University Press, 2017).

nature of morality, virtue, emotions, spirituality, and the like that they are multifaceted, and that no one facet ever captures the whole.

It is not just that morality and virtue are multifaceted, the empirical fact is that the correlations between the various facets of morality tend to be relatively low. This is not as widely recognised as it should be, because much research has focused on a particular aspect of morality or virtue, and has not concerned itself with other facets. Derek Wright's book, *The Psychology of Moral Behaviour*,[11] emphasised this important message about the multifaceted nature of morality. The core message was well conveyed by the cover of the book, which had a pentagon shape, with moral insight, resistance to temptation, altruism, belief, and guilt on the five sides of the pentagon, with the face of a child in the middle.

There is often little correlation between the various aspects of morality. For example, there is little connection between the guilt that people feel over transgressions and their actual moral behaviour. People can be wracked with guilt about behaving badly, but still go on behaving badly. Guilty feelings don't seem to result in good behaviour. Similarly, intellectual knowledge about morality, ethics, and virtue hardly correlates at all in the general population with other aspects of moral functioning.

Particularly important for the computational modelling of virtue is the dissociation between values and behaviour. People can sincerely hold strong moral values, but not act on them when particular situations arise. There is a large literature in social psychology on the failure of bystanders to help people experiencing some kind of personal crisis, despite the fact that many of those who failed to help believed themselves to be the kind of person who would help in such situations.[12] They often believe that they actually would help; but still, in practice, they don't. Moral values and moral behaviour often don't line

[11] Derek Wright, *The Psychology of Moral Behaviour* (London: Penguin, 1971).

[12] M. Levine et al., "Identity and Emergency Intervention: How Social Group Membership and Inclusiveness of Group Boundaries Shape Helping Behaviour," *Personality and Social Psychology Bulletin* 31:4 (2005): 443–453, https://doi.org/10.1177/0146167204271651.

up together at all well (though they may do so in exceptionally virtuous people more than in the rest of us). Again, this common dissociation between moral values and moral behaviour is not as well recognised as it should be, because too much psychological research on virtue just uses self-report measures, which only picks up what people believe about themselves, not what they actually do.

Virtue is also multifaceted, as Blaine Fowers and his colleagues commented in a good recent overview of the science of virtue. As they say, "Virtues are (a) behaviourally expressed, (b) based on knowledge about the virtue, (c) accompanied by concordant motivation and emotion, and (d) expressive of a stable disposition."[13] Their theoretical model of virtue, STRIVE–4, distinguishes each of these facets and assumes that each plays an important role in virtue.

One of the most significant issues for the computational modelling of virtue is how to handle the relationship between values and behaviour in human virtue. I assume that it would be possible in principle to take an approach to the computational modelling of virtue in which values were always reflected in behaviour. It might be argued that that would be a superior approach to virtue, compared with the struggle humans have with the behavioural expression of their moral values. However, a computational approach to virtue that is human-like would need to find a way to model the difficult relationship between values and behaviour that humans exhibit.

## Relationality and Virtue

Morality and virtue are inherently relational; they are not private matters. So, if AI is to be virtuous, it will need some kind of relational intelligence, an issue discussed by Noreen Herzfeld.[14] AI research has so far taken a very individualistic approach to intelligence and has

[13] Blaine J. Fowers et al., "The Emerging Science of Virtue," *Perspectives on Psychological Science* 16:1 (2021): 118–147, https://doi.org/10.1177/1745691620924473.

[14] Noreen Herzfeld, *The Artifice of Intelligence: Divine and Human Relationship in a Robotic Age* (Minneapolis: Fortress, 2023).

hardly tried to develop relational forms of AI. However, there have been calls for AI to move in this direction. William Clocksin, in particular, has drawn attention to the fact that human intelligence is relational, and that AI needs to be relational as well if it is to simulate human intelligence.[15] In the language of 4E cognition,[16] human intelligence is socially embedded, as well as embodied, enacted, and extended. Human-like AI will also need to have these features.

The currently prevailing, highly individualistic assumptions about intelligence seem to have arisen towards the end of the 19th century, in what is sometimes called "late modernity." The development of intelligence tests was a product of individualistic assumptions, and helped to entrench them further. Previously, there had been more transpersonal assumptions about intelligence, as something in which people *participated*, rather than as something that they possessed.[17]

AI need not be as individualistic as it has been so far. Given that relationality is a core feature of human nature, progress in developing androids that have human-like intelligence depends on being able to program relationality.[18] Clocksin has recently taken practical steps towards computational modelling of friendship, focusing on caregiving as a core feature of friendship, and using Affinity Modelling.[19] It is a

[15] William F. Clocksin, "Artificial Intelligence and Human Identity," in *Consciousness and Human Identity*, ed. J. Cornwell (Cambridge: Cambridge University Press, 1988); William F. Clocksin, "Artificial Intelligence and the Future," *Philosophical Transactions of the Royal Society A* 361: 1721–1748. Reprinted in *Society, Ethics, and Technology*, ed. M. Winston and R. Edelbach, Fourth edition (London: Wadsworth, 2003).

[16] A. Newen et al., *Oxford Handbook of 4E Cognition* (Oxford: Oxford University Press, 2018).

[17] Harris Wiseman and Fraser Watts, "Spiritual Intelligence: Participating with Heart, Mind, and Body," *Zygon: Journal of Religion and Science* 57:3 (2022): 710–718, https://doi.org/10.1111/zygo.12804. Fraser Watts and Marius Dorobantu, "The Relational Turn in Understanding Personhood: Psychological, Theological, and Computational Perspectives," *Zygon: Journal of Religion and Science* 58:4 (2023): 1029–1044, https://doi.org/10.1111/zygo.12922.

[18] William F. Clocksin, "Steps toward Android Intelligence," in *The Cambridge Companion to Religion and Artificial Intelligence*, ed. Beth Singler and Fraser Watts (New York: Cambridge University Press, in press).

[19] William F. Clocksin, "The Affinity Program: Computer Program," 2022, available with supplementary material at https://www.issr.org.uk/projects/

significant development in computational AI. Caregiving raises moral issues, and one of the marks of virtue in a robot would be its capacity for caregiving.

There is much further to go before we will have the kind of relational intelligence that is necessary if AI is to be virtuous. Steps will need to be taken, for example, to integrate relational intelligence with the kind of monitoring of virtuous behaviour that is an important feature of a virtuous person. Monitoring of virtue requires, among other things, a degree of empathy for how the other person is responding to one's actions.

## Process Rather Than Content

It is worth noting that not all areas of moral psychology are equally useful in computational modelling of human virtue. I am focusing here on *process*, rather than content. For example, one of the most interesting recent developments in understanding the moral mind is Jonathan Haidt's approach to the various different themes that govern the moral thinking of different people.[20] However, he focuses on content rather than process, whereas a computational implementation of virtue will need to focus on process.

I have also chosen not to work with the well-known approach to stages of moral development proposed by Lawrence Kohlberg.[21] There is useful material there, especially in Kohlberg's later work. However, Kohlberg's approach to moral development, like James Fowler's rather similar approach to faith development, suffers from being a conflation of various items.[22] It synthesises different elements such as (i) the shift from concrete to abstract thinking, well known from the work

understanding-spiritual-intelligence/. William F. Clocksin, *Computational Modelling of Robot Personhood and Relationality* (Berlin: Springer, 2023).

[20] Jonathan Haidt, *The Righteous Mind: Why Good People are Divided by Politics and Religion* (New York: Pantheon, 2012).

[21] Lawrence Kohlberg, *Essays on Moral Development: Vol. II. The Psychology of Moral Development: The Nature and Validity of Moral Stages* (San Francisco: Harper & Row, 1984).

[22] Watts, *Psychology, Religion and Spirituality*.

of Piaget; (ii) a widening circle of social contexts which is a feature of most children as they grow up, and (iii) an element of ideology about the later stages of moral development. There are too many different things going on here for it to be an approach that lends itself to computational implementation. I am also doubtful as to whether talking of "stages" is the right approach, as it seems that earlier "stages" are not replaced by later ones, but coexist with them.

## Interacting Cognitive Subsystems (ICS)

I will now try to develop a systematic approach to this problem of the fragmentation of human virtue, and connect it with a cognitive architecture that can be employed in preparing the ground for full computational implementation, the Interacting Cognitive Subsystems developed by Philip Barnard.[23] But first I will make some general points about the value of cognitive architectures. They are a long-standing hybrid between cognitive psychology and AI, and therein lies their value. The interface with empirical research in cognitive psychology keeps cognitive architectures grounded in how humans do things, such as being virtuous. This gives them a better chance of modelling virtue in a way that is not just human-level but human-like.[24] In addition, they are also looking towards computational implementation, which imposes a greater requirement for precision than is often found in psychological theorising. I would claim that cognitive architectures benefit both psychology and AI, but in different ways.

I believe that cognitive modelling of virtue in terms of Interacting Cognitive Subsystems would be a significant step forward. Recent work by Anthony Ahrens and David Cloutier[25] has developed an engagement

[23] Teasdale and Barnard, *Affect, Cognition and Change*; John D. Teasdale, *What Happens in Mindfulness: Inner Awakening and Embodied Cognition* (New York: Guilford Press, 2022).

[24] Marius Dorobantu, "Human-Level, but Non-Humanlike: Artificial Intelligence and a Multi-Level Relational Interpretation of the Imago Dei," *Philosophy, Theology and the Sciences* 8:1 (2021): 81–107, https://doi.org/10.1628/ptsc-2021-0006.

[25] Anthony Ahrens and David Cloutier, "Acting for Good Reasons: Integrating

between Catholic virtue theory and cognitive psychological models, but has apparently not yet engaged with any particular cognitive architecture that has computational-level precision. Also relevant is work on implementing virtue in robots, such as that of David Crook and Joseph Corneli.[26] However, I think there is value in working first at the level of cognitive architectures, which stay closer to cognitive psychological theories, before moving to computational implementation.

In cognitive psychology it is important to distinguish between two different modes of central cognition, as humans have at their disposal two very different modes of cognition, as many theorists have proposed, albeit using different terminology.[27] One is intuitive, embodied, affective, and holistic; what Barnard calls the "implicational" subsystem. It is the mode of cognition that is similar to, and developed from, the central cognition of our primate ancestors. However, humans also have a more conceptual mode of cognition, which is often linguistic, but not necessarily so; what Barnard calls the "propositional" subsystem. It is more detached, less participatory, often working with representations of sensory experience, rather than with sensory experience itself. It involves a slower kind of cognitive processing, though not in exactly the same way as Daniel Kahneman distinguishes fast and slow processing, as explained in detail by Harris Wiseman.[28]

Virtue Theory and Social Cognitive Theory," Social and Personality Psychology Compass 13:4, (2019): e12444, https://doi.org/10.1111/spc3.12444. David Cloutier and Anthony Ahrens, "Catholic Moral Theology and the Virtues: Integrating Psychology in Models of Moral Agency," Theological Studies 81:2 (2020): 326–347, https://doi.org/10.1177/0040563920928563.

[26] David Crook and Joseph Corneli, "The Anatomy of Moral Agency: A Theological and Neuroscience Inspired Model of Virtue Ethics," *Cognitive Computation and Systems* 3:2 (2021): 109–122, https://doi.org/10.1049/ccs2.12024.

[27] Fraser Watts, "Dual System Theories of Religious Cognition," in *Head and Heart: Perspectives from Religion and Psychology*, ed. Fraser Watts and Geoff Dumbreck (West Conshohoken, PA: Templeton Press, 2013). Marius Dorobantu and Fraser Watts, "Spiritual Intelligence: Processing *Different* Information or Processing Information *Differently*?" *Zygon: Journal of Religion and Science* 58:3 (2023): 732–748, esp. 736, https://doi.org/10.1111/zygo.12884.

[28] Harris Wiseman, "Knowing Slowly: Unfolding the Depths of Meaning," *Zygon: Journal of Religion and Science* 57:3 (2022): 719–743, https://doi.org/10.1111/zygo.12808.

I maintain that there is a major evolutionary transition between humans and other higher primates, and that humans are unique in having two different modes of central cognition. It is not that humans are "better" in any general sense; but that, for better or worse, they are different from other species. Having both of these modes of cognition available gives humans considerable cognitive versatility. The downside is that the two modes, roughly "heart" and "head,"[29] can pull people in opposite directions, leaving the person confused and not knowing what they really think or want to do. It can also lead to inconsistencies and fragmentation, in virtue and in other matters. People often have conceptualisations about what they are up to that are at variance with what is actually going on below the conceptual level. I suggest that lack of coordination between the two modes of central cognition provides a theoretical framework for understanding the discrepancy that often occurs between values and behaviour in virtue.

Barnard proposes that humans have a nine-subsystem architecture, built around these two central subsystems. The four-subsystem architecture of animals such as a zebra evolved into the nine-subsystem architecture of humans.[30] The cognitive demands that led to those developments, and the new capacities that resulted from the addition of extra subsystems, have been specified. In addition to the two central subsystems there are also three peripheral sensory subsystems (visual, auditory, and body state), two effector subsystems (limb and articulation); and two intermediate subsystems (one linking auditory and articulatory, and supporting verbal imagery; the other linking visual and limb, and supporting visuo-spatial imagery). There is no controlling central homunculus. Each subsystem uses a different code, and much cognitive work is done by information being transferred from one subsystem to another, and being recoded in the process. ICS has not yet been fully implemented, but various lines of work have

[29] Watts and Dumbreck, eds, *Head and Heart*.

[30] Philip J. Barnard et al., "Toward a Richer Theoretical Scaffolding for Interpreting Archaeological Evidence Concerning Cognitive Evolution," in *Cognitive Models in Palaeolithic Archaeology,* ed. T. Wynn and F. Coolidge (Oxford: Oxford University Press, 2016), 45–67.

been undertaken on partial implementation that indicate that computational implementation is achievable.[31] It is specified sufficiently precisely that it can already be regarded as a computational-level cognitive architecture.

Why use this particular cognitive architecture, Interacting Cognitive Subsystems? Much computational modelling of the human mind has focused on micro-cognitive systems, rather than being the kind of comprehensive cognitive architecture that is needed for modelling something such as virtue. ICS arose from empirical work on psycholinguistics but, as one would expect of a general cognitive architecture, it has been applied to a wide range of cognitive functioning including human-computer interaction, depression, mindfulness and other spiritual practices, and dance.[32] Its track record suggests that it can be applied to the modelling of human virtue. The ICS distinction between different modes of central cognition promises to be fruitful.

In religion and spirituality there is often a dissociation between conceptualisations and experience. Anthropologists such as Robin Dunbar often make a distinction in the evolution of religion between shamanic religion and the later doctrinal religion that developed when there were fixed settlements.[33] I have suggested elsewhere that this maps onto the distinction between conceptual and intuitive-holistic cognition in ICS. The cognition associated with trance dancing seems to have been largely of the latter kind, but religious cognition became much more conceptual in the doctrinal phase.[34]

[31] Fraser Watts, "Cognitive Modelling of Spiritual Practices," in *The Cambridge Companion to Religion and Artificial Intelligence*, ed. Beth Singler and Fraser Watts (New York: Cambridge University Press, in press).

[32] Philip J. Barnard, "Paying Attention to Spiritual Meanings: A Manifesto for the Cognitive Modelling of Contemplative Practices," International Society for Science and Religion, 2023, https://www.issr.org.uk/wp-content/uploads/2023/05/Paying-Attention-to-Spiritual-MeaningsPJB2023.pdf.

[33] Robin Dunbar, *How Religion Evolved and Why it Endures* (Oxford: Oxford University Press, 2022). See also Fraser Watts and Marius Dorobantu, "Shamanic and Doctrinal: Dunbar and the Spiritual Turn in Contemporary Religion," *Religion, Brain & Behavior* 14:1 (2024): 85–90, https://doi.org/10.1080/2153599X.2023.2168733.

[34] Fraser Watts, "The Evolution of Religious Cognition," *Archive for the Psychology of Religion* 42:1 (2020): 89–100, https://doi.org/10.1177/0084672420909479.

ICS also helps to make sense of the current development of people who regard themselves as "spiritual but not religious."[35] Mainline religion, at least in the Western hemisphere, is now widely felt to be over-conceptual and insufficiently experiential. The current widespread interest in spiritual practices such as mindfulness, rather than religion, can be understood as an attempt to get back to something more experiential.[36] Spiritual practices such as mindfulness have been modelled, using interactive cognitive subsystems, as involving the prioritisation of intuitive-holistic cognition over conceptual cognition.[37] The same is true of Christian spiritual practices such as the Jesus prayer.[38]

## Conceptual and Intuitive-Embodied Aspects of Virtue

The central problem about virtue, framed psychologically, is how to deliver virtuous behaviour. That is the question that needs to be understood psychologically, and which needs to guide the computational implementation of human-like virtue. Believing in virtue, having virtuous intentions, or having warm empathic feelings, don't constitute virtue unless there is actual virtuous behaviour.

I suggest that this problem is helpfully approached in terms of the distinction that ICS makes between the two modes of central cognition, one intuitive-embodied, the other conceptual. There are routes to behaviour from either central subsystem. However, I suggest that neither central subsystem alone can deliver virtuous behaviour on a consistent basis. My theoretical proposal here is that consistently virtuous behaviour depends on the coordinated activity of both central subsystems. I suggest that is what is involved in the education of conscience, i.e., developing a sustained coordination between the two central subsystems to deliver consistently virtuous behaviour.

[35] Galen Watts, *The Spiritual Turn: The Religion of the Heart and the Making of Romantic Liberal Modernity* (Oxford: Oxford University Press, 2022).

[36] Watts, "The evolution of Religious Cognition."

[37] Teasdale, *What Happens in Mindfulness*.

[38] Watts, "Cognitive Modelling."

It is not hard to see the limitations of either central subsystem on its own. Considering first the limitations of conceptual cognition, it is all too easy for good intentions framed at the conceptual level not to result in the desired behaviour, as everyone knows who has made New Year resolutions. There is also huge scope for human self-deception, and for people to believe that they are more virtuous than is evident in their actual behaviour, as is illustrated by the research on help by bystanders, already noted above.

There is also the consideration that many decisions have to be made intuitively and instantaneously. Humans rely more on intuitive-embodied cognition when under stress, or when they lack time or cognitive energy for another reason. For example, soldiers on the battlefield, have no time to think through at a conceptual level what would be the virtuous course of action. Conceptual-level virtuous intentions need to be supported by the more intuitive-embodied level of cognition if they are to result in virtuous behaviour.

However, on the other hand, there will not be, and cannot be, any real movement towards virtue without the involvement of conceptual-level cognition. It seems that the language-based conceptual ability of humans gives them a distinctive capacity to make moral commitments. Humans have a distinctive capacity to *decide* to do good (or evil), though they may not always carry out their decision successfully. It is perhaps that capacity to make conscious moral decisions that is being referred to in the story of the Garden of Eden in Genesis 3 as the "knowledge of good and evil."[39]

There is something deliberate and intentional about virtue, that goes beyond simply behaving in a way that serves the interests of other creatures. Other species can behave in ways that benefit other creatures, but I suggest that it would be stretching the concept of "virtue" to suggest that non-humans were virtuous. I suggest both conceptual cognition and intuitive-embodied cognition contribute to intentional behaviour (including virtue), but in different ways.

[39] Fraser Watts, *Theology and Psychology* (Basingstoke: Ashgate, 2002).

My proposal is that the education of conscience starts with a commitment at the conceptual level of cognition. However, in order for that to result in sustained and consistent virtuous behaviour, the conceptual commitment needs to be transferred to the intuitive-embodied level of cognition. That is possible, but it is a long, slow process. It happens through a series of events, just as the initial acquisition of morality in infants depends on internalisation through a series of specific events. Something similar is envisaged in the ecological approach to socio-emotional learning developed by Stephanie Jones and colleagues.[40] There is also an element of this in the approaches to moral intelligence of Doug Lennick and Fred Kiel,[41] or Michele Borba.[42]

## Moral Mindfulness

If virtue is to be acquired (or if conscience is to be educated), people need to monitor their virtuous behaviour in an open-minded and honest way. It involves people actually noticing how virtuous they are being (or not), more than is usually the case. In that regard, it is akin to mindfulness. In learning a mindful mode of attention people often attend to their bodies, perhaps breathing, or the soles of the feet. There are good reasons why, in learning mindfulness, it is a good strategy to focus on the body.[43] However, mindful attention can in principle be applied to anything, including virtue.

What I am proposing here is a kind of moral mindfulness, which involves observing in a sustained way how virtuous one is being. Most people are probably not very observant about their virtue, but they can

[40] R. Bailey et al., “Getting Developmental Science Back into Schools: Can What We Know about Self-Regulation Help Change How We Think About ‘No Excuses’?” *Frontiers in Psychology* 10 (2019): 1885, https://doi.org/10.3389/fpsyg.2019.01885.

[41] Doug Lennick and Fred Kiel, *Moral Intelligence: Enhancing Business Performance and Leadership Success* (New York: Prentice-Hall, 2005).

[42] Michele Borba, *Building Moral Intelligence: The Seven Essential Virtues that Teach Kids to Do the Right Thing* (New York: Jossey Bass, 2002).

[43] Fraser Watts, *A Plea for Embodied Spirituality: The Role of the Body in Religion* (Norwich: SCM Press, 2021).

choose to make it a focus of attention. It is similar to how most people are not very observant of their dreams, but they can choose to keep a dream diary, which will make them more aware of their dreams. Monitoring of virtue occurs in different ways at the conceptual and intuitive-embodied levels, and both can make a useful contribution to the development of virtue. Su et al., using ICS, made a distinction between "glance" and "look," which seems a promising distinction in the modelling of human cognition.[44] Conceptual monitoring of virtue will probably involve "looking," whereas intuitive monitoring may just involve a "glance."

It is also necessary for a disciplined commitment to observing one's virtue, or lack of it, to be carried out in an honest and undistorted way. That is easier said than done, as it seems to be very easy for humans to believe the best about themselves or, in the case of depression, to go to the other extreme and believe the worst about themselves. Research on "depressive realism" has shown that most people have a rosy glow in how they perceive themselves, for example believing that they are more highly regarded by other people they actually are.[45]

Iris Murdoch who, in her philosophical writings has been one of the main advocates of a virtue approach to ethics,[46] has provided in her novels a series of instructive case studies of how people mess up their own lives, and the lives of other people, by perceiving what is going on in a distorted and self-serving way. A. S. Byatt in *Degrees of Freedom*[47] provided a helpful commentary of that theme in Murdoch's novels, by bringing the narrative stories of the novels into dialogue with Murdoch's theoretical writings. As Murdoch sees it, learning to be virtuous involves unlearning self-serving and distorting perceptual

[44] L. Su et al., "Glancing and Then Looking: On the Role of Body, Affect, and Meaning in Cognitive Control," *Frontiers in Psychology* 2 (2011): 348, https://doi.org/10.3389/fpsyg.2011.00348.

[45] L. Y. Alloy and L. Y. Abramson, "Depressive Realism: Four Theoretical Perspectives," in *Cognitive Processes in Depression,* ed. L. B. Alloy (New York: Guilford Press, 1988), 223–265.

[46] Iris Murdoch, *Metaphysics as a Guide to Morals* (London: Chatto & Windus, 1992).

[47] A. S. Byatt, *Degrees of Freedom: The Early Novels of Iris Murdoch* (London: Vintage, 1994).

habits. Murdoch's approach is much indebted to the spiritual writings of Simone Weil.[48]

## Conclusion

I am conscious of having only sketched out, in a very preliminary way, a potential computational approach to the acquisition of virtue. However, through framing my proposal in terms of a computational-level cognitive architecture (ICS), I hope I have pointed the way to how there might be computational implementation of human-like virtue. Some lines of work in moral psychology are more helpful than others in the cognitive modelling of virtue.

My focus here has been on how we might develop a computational theory of the acquisition of virtue, with all the rigour and precision that would bring to understanding this topic of human significance. It would be very challenging work to carry through, but I have suggested how it might be approached. My focus has been on computational theorising about *human* virtue, rather than on developing a robot with some other kind of virtue that was not human-like. However, in the long run, the approach which I have suggested here might subsequently help in actually being able to develop human-like virtue in a robot.

*Acknowledgment:*

This work was supported by a grant from the John Templeton Foundation to The Center for Theology and the Natural Sciences (CTNS), for a project on "Virtuous AI? Artificial Intelligence, Cultural Evolution, and Virtue.

*The author reports there are no competing interests to declare.*

Received:  05/03/24. Accepted: 05/06/24  Published: 09/09/24

[48] Silvia Caprioglio Panizza, *The Ethics of Attention: Engaging the Real with Iris Murdoch and Simone Weil* (London: Routledge, 2022).