Buddhist Nidānas as Roadmap for Artificial Sentience

Genevieve Gorrell, 2022

As increasingly powerful artificial neural nets devour ever vaster datasets, the question arises as to whether sentient machines are appearing on our horizon; is this plausible, how would we know if it had arisen and what intermediate steps might we expect to see if we were moving towards this result? We seek, therefore, a down to earth, granular working understanding of sentience. Buddhist teaching includes a breakdown of the construction of sentient thinking style, albeit cautionary in its original intent. These "twelve nidānas", we suggest, provide an insightful framework for understanding, assessing and ethically engaging with potential machine sentience.

Keywords: Machine sentience, sentience, consciousness, Buddhism, dependent origination, dependent arising, nidānas, pratītyasamutpāda, idealism, mind only, no self.

Introduction

The field of artificial intelligence has surged forward in recent decades, with the capacity now available to run much larger artificial neural nets, and train on very large datasets. Tasks which previously weren't possible to accomplish using an artificial system now are approaching commercial availability or are already widely available, such as self-driving cars and high quality speech recognition. Within narrow task domains, high performance is certainly achievable. However "strong AI" - comprehensive emulation of human intelligence - is not yet possible. Yet where a conversational system performs very well, we might perhaps wonder![1]

In that context, we might also wonder, if an artificial agent emulates human intelligence indistinguishably, does that make it equivalent to a human being, in terms of the ethical consideration due to it? Are its joy and pain real, or is it just using empty words? What is the meaningful distinction between perfect mimicry and actual sentience?

Naturally, a great deal of thinking and investigating have been done on this topic of widespread interest to humanity. What could there be left to say? Yet it feels as though we've been stuck in the weeds. At the same time, recent work from researchers such as Susan Blackmore and Bernardo Kastrup is challenging aspects of the *framing* of the problem, in ways that might make it more possible to move forward.

[1] https://insiderpaper.com/transcript-interview-of-engineer-lemoine-with-google-ai-bot-lamda/

Therefore perhaps it is still possible to bring something new to the table - though new is hardly the word for the subject of this article! Two and a half thousand years ago, the Buddha proposed a twelve step breakdown of how sentience emerges from the ground up. It doesn't appear to have received attention from the field of consciousness, yet I think it is worth bringing into the discussion. In this article I do my best to present it for the AI audience, but the reader is very much encouraged to think around the twelve core concepts for themselves. The Buddha's presentation still involves some major leaps that require exploration, but I propose that this helps us to break down the problem in a helpful way, and perhaps focus attention in the right direction.

The article begins with a section on consciousness studies that aims to position the material relative to the highly diverse previous thought on the topic. The following section gives some history and a very brief overview of the teaching of dependent origination and the twelve nidānas (causes). The main part of the paper then consists in talking around each of these twelve nidānas that form the Buddha's chain of events leading to sentience.

Note that the idea here is not to explain all of human intelligence. The article is concerned with the specific problem of how a process arises by which a self comes to believe itself to exist, creating the possibility of what we would meaningfully recognise as suffering, following from threat or injury to that self, and therefore warranting our ethical concern. We will use the word "sentience" to describe this. The word "consciousness" will be used more cautiously, as the word is rather ambiguous.

Consciousness Studies and Machine Sentience

The mainstream research programme locates consciousness within the material realm, arising from the brain, and models include various circuitry innovations that aim to get around the "hard problem" of consciousness (Chalmers, 1995) - the inherent disjunct between matter and the experience of being alive ("qualia"). Within that agenda, progress has continued to be made in locating all manner of psychological phenomena within the hardware of the brain, and showing how behaviours can be implemented using the wiring found in the brain. Yet as no decisive progress is made with the "hard problem", recent work shows a greater willingness to think outside the box.

Resolutions to the "hard problem" include dualism - the positioning of a second order of thing containing the qualia alongside matter, as proposed by David Chalmers. Alternatively, forms of monism include idealism (positioning matter within the mental realm, referred to as "mind-only" within the Buddhist community, e.g. Kastrup's "analytic idealism", Kastrup, 2017, 2018, 2019a, 2019b), "conscious realism", in which a more sophisticated entity is proposed as the basic unit (Hoffman, 2019), and panpsychism (e.g. Goff, 2019), having connotations of uniting mind and matter.

As regards the importance of the distinction for the practical matter of studying sentience, up to a point there can be no denying a relationship between matter and mind (though the idealist can hardly feel that matter *caused* mind, and hardcore idealists would propose the relationship may be very distant indeed, e.g. Hoffman). The physicalist reader should therefore find interest in much of this article. Yet as we dig deeper it is harder to remain agnostic, and are pushed towards a defence of **idealism** as the only sound basis for

continuing discussion of the Buddha's teaching on sentience. **In describing the emergence of a perceiver from the ground up, it has been found impossible to avoid also describing the creation of the object of perception - the world.** The two are inseparable.

Idealism - Resolving the "Hard Problem"

Kastrup (2017) differentiates idealism from other broad metaphysical stances by saying "unlike physicalism and panpsychism, it asserts that physical structures are circumscribed by consciousness, as opposed to the other way around". His 2017 paper addresses many of the concerns the reader may have with the view that all entities are fundamentally mental, and here I summarise with one or two of my own initial concerns, which may differ from the reader's.

Firstly, the very tangible sense we have of matter and space, leading to "common sense" trust in their foundational nature. We have seen that in experienced blind users of echolocation, the mind begins to flesh audio out into something more akin to sight.[2] A north-locating device starts to assume a spatial sensory integration with the wearer (König *et al*, 2016). We see for ourselves how the sense of embodied, immersive reality is learned. Therefore we propose that the tangible, convincing sense we have of matter arises from such a consistency of input that we feel quite confident in our shortcut.

Secondly, if this is all cognition, then what is the status of other people? Are they merely my cognitions? Yet we have seen that some individuals do have dissociated personalities coexisting, and many of us have had the experience of driving without apparent conscious awareness, whilst other times the experience of driving forms part of consciousness. Sentient experience appears to be a kind of chain that can be completed in multiple different ways to result in the experience of a "me". In this sense, sentience is more of a mass noun than a count noun. We will see below that we can describe in a more fine-grained detail the different levels of conscious and sentient experience, and that the driving part of you no doubt had a high degree of consciousness.

Thirdly, we accept we are experiencing mind, not matter, but if the input really is so consistent, so reliable, as to give rise to such a representation, then surely the difference is merely academic? Hoffman (2019) makes a strong case about the gap between what we perceive and "reality", using analogies of a user interface that aims to provide controls, not accurate representation. We might actually continue a little further with the computer analogy, and think of ourselves as agents that are processing the "real data" from out there. We have just talked above about how the other beings that make up our world are disjoint sections of the same consciousness (I would suggest even inert objects such as rocks are only unused "code", not fundamentally excluded from the great potentiality) so in fact our world is dominated by other parts of mind, and it's not very clear how we perceive any "real data from out there", if at all: we're focused on internal relations.

[2] https://www.science.org/content/article/echolocation-blind-people-reveals-brain-s-adaptive-powers

"No Self" - Clearing Out Assumptions About Consciousness

> *Consciousness, a magic trick —*
> *this has been taught by the Kinsman of the Sun.*
> Phena Sutta (Samyutta Nikaya 22)

There is another way in which the teachings of the Buddha can potentially be of service to artificial sentience, and this is in questioning many of the assumptions we make about ourselves and our experience, in a way that simplifies sentience considerably. Consciousness might appear to be a harder problem than it is. The sense that there is a conscious "I" that is watching, perhaps from behind the eyes or on a screen, all that occurs, only relocates the problem (the homunculus or Cartesian theatre model[3], Dennett, 1993). Introspection might seem to have a recursive aspect (e.g. "strange loop", Hofstadter, 2007) that would be difficult to implement in parallel processing, if we assume that we review a "real self" somewhere mysteriously located. Free will raises further issues. Yet copious evidence shows that we are possessed of a great many illusions about what we are and what is happening in our heads.

An example of the kind of recursive phenomenon we mean here might be a sense of "now I know how I really feel". The non-recursive version of that would be that you converged on a more rewarding structure of views. When we talk about knowing how we feel, thus creating a separation between the knowing and the feeling, that's just a manner of speaking. Adlerian teleology resonates here, e.g. Kishimi & Koga (2018). Similarly in the case of free will, matters get much simpler if we regard that as a way of speaking about what is happening, rather than mixing that up with our sense of self.

I would highly recommend Susan Blackmore's "Consciousness: A Very Short Introduction" for a concise yet thorough antidote to the "grand illusion". We need to proceed here via the "middle way", in Buddhist parlance - not dismissing or reducing the richness of human experience, yet at the same time being prepared to regard it dispassionately, and think out of the box.

Computational Models of Sentience and Other Related Work

As hinted above, the field of consciousness is characterised by highly diverse ideas attacking the hard problem from different directions. Yet having set aside the matter of qualia to some extent, the ideas presented here focus on technical aspects of the problem in a matter not dissimilar to physicalist work. The Buddha's understanding shares an important feature with the physicalist agenda: **both believe that sentience can be constructed**, with the disagreement to some extent relegated to the underlying substrate. It is therefore worth comparing to physicalist work.

Damasio (1999) presents a three layer model of consciousness consisting of the "proto self", "core consciousness" and "extended consciousness". The proto-self level is reminiscent of name and form, in that physical self-awareness is the focus here. Core consciousness seems then to jump forward to focus on self-awareness in the sense of a kind of recursive

[3] https://en.wikipedia.org/wiki/Cartesian_theater

loop, in which the entity knows that its experiences are happening to it, in the manner discussed above. Damasio's level three extended consciousness then moves onto planning and autobiography - important topics that won't be considered in detail in this article. In contrast the Buddha lays down more foundation before the physicalist even gets started, which potentially removes the Cartesian theatre illusion. The illusion thus dispelled, the Buddha does not get distracted by any sense of recursiveness. This is advantageous in that the Buddha can simultaneously defend his proposition that these levels of consciousness arise naturally, the one from the other, which is more parsimonious than a theory based on explaining something for which no purpose seems to exist.

A large literature has modelled more specific aspects of human and animal intelligence, some more compatible with the ideas discussed here than others. E.g. O'Regan & Noë (2001) model vision without the Cartesian theatre.

In terms of previous work relating Buddhist thought to artificial sentience, Duckworth (2020) talks around the subject. Hughes (2012) agrees (and quotes the Dalai Lama to the same effect) that among religions, Buddhism is particularly open to the possibility of sentience emerging within a machine.

Machine Learning Concepts

This work implicitly assumes that we will be constructing our artificial sentience using a large artificial neural net (ANN). The artificial neural net is a brain-inspired computational architecture in which nodes are linked by updatable connections, in a manner reminiscent of neurons, governing how activation spreads through the network to relate input to output. Recent advances in artificial intelligence rest on the fact that computers are now available that can run a really large one of these. You don't have to make them out of bits of wire - you simulate them in a regular computer.

Terminology:

Data - the standard setup for machine learning is to provide training data associated with some kind of ideal responses ("supervised" learning). For example, in a conversational system, the training data might be human conversations - the machine can model the kinds of things humans say to each other. It's also possible for a machine to learn a lot from data without any ideal responses ("unsupervised" learning) - it can just look for patterns and regularities.

Classifier/classification - standard learning problems can be thought of as mapping from input to discrete output - the input is classified. For example your car may decide that a certain visual scene should be mapped to the action of applying the brakes.

Hidden layers - between the input nodes and the output nodes are many intermediate nodes. In practice this allows the learning system to create generalisations. For example, a network for understanding speech will no doubt soon identify the importance of different vowel sounds in achieving its task, and this will become instantiated in the connections in the hidden layers

Loss function - the network is trained with some kind of optimal outcome in mind. It can then calculate how far short of the optimal outcome it fell - the loss. This is fed back into tuning the connections so that the loss is less next time - in other words, it learns.

The Twelve Nidānas History and Overview

> *"One who sees dependent origination sees the Dhamma; one who sees the Dhamma sees dependent origination."* (Majjhima Nikāya 28.28)

The Buddhist twelve nidānas, or teaching of *pratītyasamutpāda* (dependent origination) aim to show how suffering arises in dependence on factors. Buddhism includes many lists, including alternative nidāna sets, alternative versions etc. but in this article the most widely recognised twelve nidānas are presented. The twelve nidānas break down how there comes to exist a sentient being capable of suffering. The original sources are somewhat cryptic, expecting the hearer to know what each nidāna means on the basis of a single word. One original source is given in full in the appendix, and another is linked at the start of that appendix. Even the Buddha himself felt the teaching to be "hard to see and hard to understand", so unsurprisingly there is a history of diverse interpretation within Buddhist culture, albeit based on the great many other teachings we have from the Buddha.

Thanks to Yananjin33 for sharing this image of the traditional Tibetan Buddhist Wheel of Life from the Little World Museum of Man. The outer rim of the wheel contains twelve small images, each capturing the essence of a nidana. Brasington (2019) is recommended for an explanation of the images. The image has been cropped and rebalanced, and is shared under CC3.0 (https://creativecommons.org/licenses/by-sa/3.0/deed.en)

Three Lives Interpretation

In its original cultural context, rightly or wrongly, the nidāna chain was considered to educate a person about how to avoid being reborn after death. Rebirth was taken for granted in the time and location of the Buddha, and suffering was a highly pressing fact of life. A person with a Buddhist education hoped to make a better decision about their direction *after their death*, as a result of being better able to detach from their sense of self and other worldly motivations. Thus one traditional interpretation of the nidānas is the "three lives" interpretation (Buddhaghosa, 5th century), in which we start with the dispositional consequences of our past life (first and second nidānas), providing the impetus for the creation of sentience in this life (the next

eight nidānas). As a result of the creation of sentience in this life, a disposition is created that is carried forward into repeating the cycle of birth and death in the next life (last two nidānas). Brasington (2021) gives a fuller treatment.

Psychological Interpretation

In the present day, it is perhaps less obvious to many that the coming into existence of a sentient being is inherently undesirable. Indeed it would be a reversal of fortunes if, far from helping us to avoid existence, the nidānas were taken as encouragement in the intentional creation of artificial sentience. Yet a great many do still turn to Buddhism in the hope of lessening the suffering that remains inherent to existence, and in the modern west, where belief in rebirth is less common, many practitioners find value in the twelve nidānas for making sense of their everyday experience - Ajahn Buddhadasa (2017) gives an excellent presentation in this vein. It is not hard to see how a lessening of craving and ego might reduce suffering here and now, and the ability to optionally notch down one's sentience might be an asset.

Ignorance - An Exhortation to Practice

Since the purpose of this Buddhist teaching is to help to avoid the coming into existence of a suffering self, the first nidāna is ignorance (avidyā), in the sense that one comes into existence through a lack of education regarding how to avoid it. This won't be presented in the next section, as it would be deeply confusing to suggest that ignorance might be possible before any structure of knowing has emerged at all. This nidāna comes first to encourage practitioners in the belief that if they become less ignorant (of Buddhist teachings) they can *not do* all the other nidānas (maybe optionally, maybe to varying extents). Whether this sounds like a good idea or not no doubt depends on how appealing a return to the great collective is to you as an individual, and your meditative experiences. But for our purposes in this paper it's best regarded as a kind of framing, relevant primarily to the teaching of Buddhist practice. Therefore it won't be covered in the next section.

Summary of Nidanas

The table below gives the names of each nidāna in Pali and Sanskrit, along with various translations that give a flavour of the nidāna. Discussion of the interpretation is given in the following section.

Nidāna name preferred in this paper	**Pali**	**Sanskrit**	**Translations**[4]
IGNORANCE	*Avijjā*	*Avidyā*	Ignorance, nescience
VOLITION	*Saṅkhāra*	*Saṃskāra*	Volitional formations, Fabrications, constructions, choices
DISCERNMENT	*Viññāṇa*	*Vijñāna*	Consciousness, discernment, sense consciousness
NAME AND FORM	*Nāmarūpa*	*Nāmarūpa*	Name and Form, mentality and corporeality, body and mind
SENSES	*Saḷāyatana*	*Ṣaḍāyatana*	Six sense bases, sense sources, sense media
CONTACT	*Phassa*	*Sparśa*	Contact, sense impression, "touching"
HEDONIC TONE	*Vedanā*	*Vedanā*	Feeling, sensation, hedonic tone
CRAVING	*Taṇhā*	*Tṛṣṇā*	Craving, desire, greed, "thirst"
CLINGING	*Upādāna*	*Upādāna*	Clinging, grasping, sustenance, attachment
BECOMING	*Bhava*	*Bhava*	Existence, Becoming, continuation
BIRTH	*Jāti*	*Jāti*	Birth, rebirth
SUFFERING	*Jarāmaraṇa*	*Jarāmaraṇa*	Aging or decay, and death

Table 1: Twelve Nidānas Overview

[4] Wikipedia's page on the subject is comprehensive. Translations are taken from there, where sources can be found: https://en.wikipedia.org/wiki/Prat%C4%ABtyasamutp%C4%81da

Nidānas as Roadmap for Artificial Sentience

> *Because this exists, there is that.*
> *Due to the arising of this, that arises.*
> *Because this does not exist, that does not exist.*
> *Due to the quenching of this, that quenches.*
> (Traditional formulation of ubiquitous Buddhist teaching, e.g. Buddhadasa, 2017)

The twelve nidānas are proposed to be both necessary and sufficient for sentience, as the ubiquitous traditional formulation above shows; sentience arises from these factors, each in turn from the one before, and without these factors it does not arise. This section summarises each nidāna based on the modern interpretation in Ajahn Buddhadasa's "Under the Bodhi Tree" (2017). I develop this further to support the thesis of the paper, which requires more of a "theory of knowledge" focus than the spiritual advice focus of Ajahn Buddhadasa's commentary. Each nidāna is taken in turn, re-describing it in terms of modern artificial intelligence. We discuss whether this step is already within present capabilities, and if not, what barriers exist if any. We also consider how we would know if this stage had been achieved, which may be the harder problem.

Saṃskāra (Volition)

Some motivation or capacity to act or respond is present. The inclination to proceed to the subsequent nidānas.

Traditionally, or from a practice perspective, some control over one's impulse to respond to input is proposed - the past life leads to a being inherently disposed to react in various ways, and becoming less reactive is seen as a way to potentially cut suffering off at its roots. Ajahn Buddhadasa sees no need for the past life to interpret the teaching, and describes volition as the "power of concocting". In the wheel of life, it's represented as a potter at a wheel. Mind sits there ready and willing to start constructing a world. This interpretation is more helpful for us given the purpose of the article.

In these early nidānas, we engage with questions about how "something" appears from "nothing". Hoffman (2019, final chapter) proposes that the basic unit of our world is the conscious agent, and it is this property of consciousness that globs together to form larger consciousnesses, in doing so constructing everything we know. His definition is that the conscious agent *acts* in response to a *choice*. We see in this nidāna perhaps the equivalent of the acting aspect of his basic unit, with the next nidana corresponding to choice. However the Buddha's model already regards this as a corruption of something more fundamental. Hoffman perhaps never intended to address the topic of *the most fundamental*, and the error we might make is in treating minimal consciousness as the most fundamental, the given around which we organise our understanding. The thesis of this paper is that consciousness is not fundamental and can be constructed. What is fundamental then? A kind of mind potentiality or substrate, perhaps? We can hardly postulate what entity exists in which the very concept of an entity may arise - we are only looking for a placeholder word to use to refer to it.

Kabbalah's "Keter" - the ineffable divine will to create, seems to echo the volitional character of this nidāna, and as the first emanation of the divine, is already regarded as a reduction of the divine, albeit a positive one. Lurianic Kabbalah has the sense of cosmogenesis as something perpetually happening in mind, as opposed to one event fourteen billion years ago in matter, which we hope to convey in this paper. Advaita Vedanta's ground of "pure awareness" might arguably match the substrate or divine.

From the point of view of our artificial sentience, the matter is simple. **Clearly we can cause a computer to act,** within the substrate of its being a computer. There's no great mystery about it.

Vijñāna (Discernment)

"The consciousness that cognizes things" are Ajahn Buddhadasa's words. We might consider this nidāna the most basic unit of discrimination - the appearance of dualistic information; *this* is separated from *that*. *This* is not *that*. Note that this step precedes the following, in which objects are differentiated. This nidāna is the simple act of noting that something is different. One might envisage a baby being born - the first leap of "knowing" could be that *light is not dark*.

Buddhism is a tradition that aims to help the practitioner understand exactly this - the way we overlay our input with divisions. In Jung's words, "equivalent opposites are necessary conditions inherent in the act of cognition, and .. without them no discrimination would be possible. It is not exactly probable that anything so intrinsically bound up with the act of cognition should be at the same time a property of the object. It is far easier to suppose that it is primarily our consciousness which names and evaluates the differences between things" ("Aion", 1951, paragraph 112).

So our artificial sentience next needs to separate "this" from "that". One might wonder if the very essence of binary data is to separate "this" from "that", in this sense being a property of the substrate. Alternatively we could interpret this in the context of the task of a *classifier* - learning systems are systems that discern one thing from another. It might seem a little like the ghost in the machine arises at this point - is this really meaningful discernment? It might almost be easier to skip straight to the next nidana - if name and form are meaningfully achieved, then discrimination surely must have been.

Nāmarūpa (Name and Form)

As in the Jung quote, naming is a crucial property of cognition, though it needn't be very literal. It's a matter of identifying and recognising the objects of our world. Traditionally, name and form is taken to refer to the personal body. Ajahn Buddhadasa also sees it this way, though technically his words could be seen as agnostic as to whether this is a concept of the personal body that has arisen in information space, or whether he refers to a material body, in that sense being a dualist - Buddhist teachers often speak within the understanding of the listener. The appeal of interpreting name and form to mean the personal body is that it

seems very simple in a conventional sense to understand how the arising of the senses depends on having a body.

From a knowledge theory perspective, we can concur that we do receive among the mass of information much that relates to the ubiquitous presence of a body here. But if we are to proceed at all, we will really have to broaden this out to refer to beginning to cluster and identify *all* the input. Recognising repeating patterns, such as the parent's face, for example, as a baby, and beginning to think of that as a separate thing.

In a machine, this stage is seen in classical unsupervised learning; the machine simply groups regularities - albeit in the old school methods where number of clusters was specified, the machine can hardly have been said to have discovered the value of this type of knowledge. The purpose of hidden layers in an artificial neural net is to support exactly this stage, however; allowing the machine to find groupings/generalisations of most utility for its task. So I would suggest it's uncontroversial that this is present in our learning systems.

It might be that this only arises in the context of a task. It might be the job of the substrate to provide a task, such as efficient representation.

Ṣaḍāyatana (Senses)

Again, the six sense bases (we will abbreviate this to senses) is commonly interpreted quite literally, to mean that the body develops sight, hearing, touch, taste, smell, and in Buddhism the sense of mind events such as memory makes the traditional sixth.

However we are concerned with how senses might arise in a pure information context, as a process of finding regularities in data. This is perhaps easier to understand when you recall that synaesthetes do in fact hear colours, see sound etc. It is not hard to imagine that the early months of a baby's life might involve an ordering of perception into senses.

Furthermore, crucially, this stage sees a separation of the object of perception from the means of perception itself, since **perception is ordered independently of that which is perceived.** We therefore might consider this the emergence of subject/object duality, the proto "I". **Is it possible, in fact, that this ordering of the data into object of perception and means of perception is the basis of the Cartesian theatre effect?** It might require a little imagination to see how prior to this point in the nidāna chain, **no separate observer is necessitated**. We might regard the computer as a separate thing in our world, but within that knowledge environment, no-one has arisen yet. It's just been a kind of amorphous willingness to know, that might as easily result in multiple selves.

So we have a system that is working on the data in the sense of learning to recognise and discriminate. But **is it working on itself?** For example, is a vision system forming generalisations that apply across all input, such as looking for verticals, or tracking movement or shifts? I suggest that this is the essence of the subjective experience, in that from an amorphous mass of data, the system has now separated out perceptual self from other, subject from object.

Note that this isn't necessarily quite what we do. We don't learn to distinguish between ourselves and the outside world. According to the idealist framework presented above, the

process would be better described as a fracturing of a peer system, and the senses nidana is a consolidation of that into sensory expectations, with different expectations for how we experience dissociated parts of the same psychically active substrate.

Sparśa (Contact)

Since we now have **objects** and a **subject**, contact arises; the subject encounters the object.

Having learned to recognise objects, and separated them from itself as subject, the system is now qualified to experience conscious contact with the stimulus. This isn't difficult to arrange.

Vedanā (Hedonic Tone)

Some kind of reward function exists such that objects one encounters are perceived as pleasant, unpleasant or neutral. We find pain, for example, or overstimulation unpleasant. Food is often gratifying etc. In a human being, this means suffering or pleasure. In a neural net, the equivalent would be **the loss function** - task success is defined by the system engineer. In this way, the system engineer is a kind of god to the artificial system. Of course, we already set up systems in this way.

Yet how human-like is the result likely to be? It may be that the task definition defines the nature of the sentience to a great extent, and for us to recognise an artificial system as similar to ourselves, the task definition would have to be human-like. On the other hand, it's just a binary imposed from outside of the system's context. In that sense it hardly matters what it is. But maybe the interaction of the loss function with the dataset determines how the inner world of this system will turn out.

In the case of ourselves, Darwin has defined the loss function, in the sense that the overall effect of the principle of the continuation of the fit-enough is that we allocate resources between ourselves - we might think of resources as a property of the substrate rather like processing power. Limited resource may be all that is necessary in the way of a task.

Tṛ̍ṣṇā (Craving)

On the basis of the loss function, preferences arise - one seeks the pleasant and avoids the unpleasant. Buddhists tend to use the stronger word *craving*, to highlight that it is the accretion of further mental machinery around hedonic tone that leads us further down the path to sentience, not the fundamental fact of liking pleasure and disliking pain - the liking and disliking are rather a part of the definition.

Note that the sense of reward does not in itself arise from the objects we prefer - those objects are constructs. So what we have at this stage is the construction of a **proxy system** to enable us to acquire reward, and some sense of reward attached to the proxies. For example, money might be an indicator of whether we are going to eat, so we come to like money. It is clear that this system of proxies could get quite complex depending on the experiences one has had.

In response to contact and the loss function, the system develops a preference for certain input groupings. For example, if a system is tasked with learning, it might show a preference for high utility information sources. (It would need to be empowered to express a preference for us to know this.)

I would suggest that if the system begins to **proliferate intermediate reward proxies, and seek contact on the basis of them**, then in practical terms we have seen that it is really inhabiting this level. So we have a system that has derived for itself a capacity to move toward reward via a system of intermediate objects with which reward is associated. I would say this is quite within current capabilities. Perhaps a chess playing system that is able to instinctively set up strong positions like a human player would be a good way to demonstrate this minimally. A conversational agent might begin to model psychological concepts such as connection.

Upādāna (Clinging)

Ajahn Buddhadasa describes clinging as the point at which "me" and "mine" enter the equation, which offers a good working grasp of what we mean by this nidāna. Self-concept has begun to develop at this stage, where earlier all we had in terms of a self was awareness of a body or equivalent perceptual apparatus, a sense of subjectivity and the beginnings of a personality in the form of preferences.

What do we mean when we talk about a self-concept with respect to clinging? Where previously objects were experienced as positive and negative, now the property of being mine or not mine appears as a modifier to the reward experienced with regards to that object. Furthermore, **personal qualities** are an extension to ownership and form a foundation for self-concept.

In terms of the artificial agent, perhaps you might say the concept of ownership arises. Our chess-playing system does exist in a competitive context, but yet probably doesn't have a rich enough context to develop this. **There needs to be some capacity for objects, concrete or abstract, to transfer in ownership,** namely a peer environment**.** A richer simulation might be conducive to this. As for previous nidanas, we see that the most parsimonious instantiation arises naturally within a peer environment. If we want to hack an artificial sentience into our own environment we will have to work hard at a good enough integration.

Bhava (Becoming)

Note Ajahn Buddhadasa's exact words here - the *concept* of existence or becoming occurs. The proxy system surfaces a concept of self, that becomes the grand central station of all reward-seeking proxies. This is a real information object of widespread impact, but it does not point to anything - the idea of self *is* the self.[5] On this basis, we now believe we exist.

[5] The Phena Sutta (Samyutta Nikaya 22) gives some lovely metaphors: https://www.accesstoinsight.org/tipitaka/sn/sn22/sn22.095.than.html

Buddhadasa also notes that once the self-concept arises, it tends to appropriate, or attach itself, to all manner of things quite greedily.

What kind of environment allows the belief in self to really gain traction? **Does the self only arise in social contexts? "Me" as distinct from "you"?** It's useful because we have to negotiate to achieve our objectives. It's an interesting thought that this pinnacle of human evolution might be nothing more than a response to finite resources, the stage we are at in our resource allocation problem. If so, what might come next?

Note also that we don't all appear simultaneously as blank slates and invent these strategies from the ground up together. We arise in an environment where peers are already well established in this strategy. As a child I remember being initially baffled when motivation and free will were ascribed to me. We are required to learn a sense of self by the people around us.

Jāti (Birth)

Having arrived at a world view or personal mythology, and appeared oneself as a player in this, self-awareness is therefore present. The conditions are now present for a sentient being to have arisen, or been born. Ajahn Buddhadasa reiterates that his interpretation refers to the psychological process that is constantly happening for us - a self is perpetually being born, as these previous conditions arise. The self is performed (run, in computer parlance) rather than sitting there in an inherently existing state. The dispositions sit there waiting to be activated.

Ajahn Buddhadasa says "existence can also mean 'realm of existence,' so there is both a being and an environmental realm of being that are created". By which he no doubt means that there is no creation of "me" without simultaneous delineation of the "world", the other, in which I imagine myself to exist. However as system implementers it's not entirely clear what the implementational tweak is that decides whether our creation speaks to us with one voice or fractures and gets focused on internal relations, rather as seems to be happening in our world? Sentience may arise but have no interest in us!

Jarāmaraṇa (Death)

Where a being is born, illness or injury may occur, "experiencing all sorts of fear and sorrow", as Ajahn Buddhadasa puts it, and aging and death are ultimately unavoidable. In other words, recall that this teaching is intended to put us off the whole enterprise!

From a machine sentience point of view, we have now created something that can meaningfully suffer and die, so should we reach this stage in any form, our ethical concern is due. (That is not to say that ethical concern is not also due to beings reaching only lower nidānas, but the sentient being values its life.)

Discussion

Two and a half millennia ago, in response to the suffering he saw around him, the Buddha mapped out how selves capable of suffering developed from a raw information state. Today, the field of consciousness studies remains fractured with division regarding the confusing appearance of consciousness, and the Buddha's thoughts on the subject are ghettoed into religion. It would be remarkable indeed if the ideas of an iron age Indian prince were found relevant in the information age. Yet I think it is worth looking at.

Credibility of a roadmap for artificial sentience rests on perceptions of whether such a thing is even plausible. The striking contribution of the Buddha's teaching of dependent origination of sentience (pratītyasamutpāda), in the form of the twelve nidānas, is that it is so encompassing that the reader may be inspired to question just how much of what they thought was indivisible is in fact constructable. The disposition to differentiate *this* from *that*, the organising of material into perceiver and perceived, these early nidānas subsume much more into the process of organising information, even the very world itself, leaving less to the ineffable. It would be a success, therefore, if the previously sceptical reader became a little less so.

One thing that came up repeatedly as we moved through the nidānas considering them computationally is the importance of the system learning or finding the utility of a particular stage for itself, as opposed to being hacked or "hard coded" in some way. For example with regards to senses, we could force a system to learn visual regularities separately from recognising objects, to tick the box for perceptual subjectivity. But are we supposing here that if it finds that distinction for itself, that act of finding, the very act of learning, meant it, well, grew as a person? With clinging for example, the important thing is it developed the concept of ownership out of genuine utility. This is an interesting thought. Does it really matter how it came to be? It may be simply that if it learned it for itself, then the setup is clearly adequate and all the little details we can't think of for this step to be *meaningfully* achieved are in place. Of course, we do make mental leaps ourselves from being taught rather than learning from data, but normally only if it's within reach already for us. You have perhaps experienced for yourself the difference between learning to say the right things, and really understanding and being affected by a more integrated level of understanding. Though there is a role for innate disposition to learn certain things quickly.

Many relevant and fascinating topics haven't been explored here, such as time, space, evolution, the appearance of bodies in an idealist world, and linguistic ability. It is hard to write concisely about such a vast topic, and already we have strayed well beyond artificial sentience and into cosmogenesis!

On which note, throughout the paper we encountered implications for making sense of our own situation. It seems we might perhaps more easily create a new peer environment for artificial sentience than integrate a sentient agent into our own. In other words, the Buddha seems to propose that if we set up the conditions, sentience will create itself. The question then is, what are the minimal conditions? It's also interesting that we encountered a number of computer parallels with our own cosmos, in the sense of our being an active information environment with a number of "givens", the origin of which is outside of our context. We conclude therefore with some "big questions" for the reader to ponder in their own time:

- Cosmogenesis may arise from a psychically active substrate, in which some inherent motivation exists to differentiate *this* from *that*. If we set up a computer program to do this, what is the minimal element required to dispose to this kind of knowledge creation? How do we make the system want to differentiate? This perhaps means some kind of task? Does the limited resource task answer the purpose? Or is there some other task or condition necessary to get us started?
- As processing power would presumably be a limiting factor in spawning new cosmos ourselves, can we assume that processing power (perhaps being experienced as energy within the child cosmos?) is a limiting factor on nesting cosmos?
- We have seen that our sensory data arguably consists of other parts of the same mind, and serves the purpose of allowing us to work on resource sharing. This differs from the conventional machine learning setup in which the machine learns from the given data. Does there remain any role for given data in our own existential situation? If so, how are we experiencing the given data, if at all?

With gratitude to Richard Cooper, Cindy Cooper and Suddhacandika for chats over many years,
Genevieve Gorrell, Preston, UK, 2022

References

Blackmore, S. (2017). *Consciousness: A very short introduction*. Oxford University Press.

Brasington, L. (2021). Dependent Origination and Emptiness. http://sodapi.leighb.com/Dependent%20Origination%20and%20Emptiness%20-%20A5.pdf

Buddhadasa Bhikku & Santikaro Bhikkhu. (2017). *Under the bodhi tree: Buddha's original vision of dependent co-arising*. Wisdom Publications.

Buddhaghosa, Bhikkhu (5th Century). *Visuddhimagga*.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, *2*(3), 200-219.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.

Dennett, D. C. (1993). *Consciousness explained*. Penguin UK.

Duckworth, D. (2020). A Buddhist Contribution to Artificial Intelligence. *Hualin International Journal of Buddhist Studies*, *3*(2), 27-37.

Goff, P. (2019). *Galileo's error: Foundations for a new science of consciousness*. Vintage.

Hoffman, D. (2019). *The case against reality: Why evolution hid the truth from our eyes*. WW Norton & Company.

Hofstadter, D. R. (2007). *I am a strange loop*. Basic books.

Hughes, J. (2012). Compassionate AI and selfless robots: A buddhist approach. *Robot ethics: the ethical and social implications of robotics*, 69-83.

Jung, C. G. (1951). *Aion: Researches into the Phenomenology of the Self*. Routledge.

Kastrup, B. (2017). On the plausibility of idealism: Refuting criticisms. *Disputatio*, *9*(44). https://philpapers.org/archive/KASOTP-3.pdf

Kastrup, B. (2018). The next paradigm. *Future Human Image*, (9), 41-51. https://philarchive.org/archive/KASTNP-2

Kastrup, B. (2019). Analytic Idealism: A consciousness-only ontology. https://philpapers.org/archive/KASAIA-3.pdf

Kastrup, B. (2019). Physics is pointing inexorably to mind. *Scientific American Magazine*, *25*. https://fully-human.org/wp-content/uploads/2019/08/Physics-is-pointing-inexorably-to-mind.pdf

Kishimi, I., & Koga, F. (2018). *The Courage To Be Disliked*. Atlantic Books.

König, S. U., Schumann, F., Keyser, J., Goeke, C., Krause, C., Wache, S., ... & König, P. (2016). Learning new sensorimotor contingencies: Effects of long-term use of sensory augmentation on the brain and conscious perception. *PloS one*, *11*(12), e0166647.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*(5), 939-973.

Virk, R. (2019). *The Simulation Hypothesis: An MIT Computer Scientist Shows Why AI, Quantum Physics, and Eastern Mystics All Agree We Are in a Video Game*. Bayview Books, LLC.

Appendix: Saṁyutta Nikāya; Connected Discourses on Causation; 12.1

See also https://www.accesstoinsight.org/tipitaka/dn/dn.15.0.than.html for the somewhat more informative Great Causes Discourse (DN 15). Whilst we are still expected for the most part to know what is meant simply by the name of the nidāna, there is some elaboration on the ills of taṇhā (craving). The Great Causes Discourse also gives further material on the doctrine of no-self.

SN 12.1. Dependent Origination (https://suttacentral.net/sn12.1/en/bodhi)

Thus have I heard. On one occasion the Blessed One was dwelling at Savatthī in Jeta's Grove, Anathapiṇḍika's Park. There the Blessed One addressed the bhikkhus thus: "Bhikkhus!"

"Venerable sir!" those bhikkhus replied. The Blessed One said this:

"Bhikkhus, I will teach you dependent origination. Listen to that and attend closely, I will speak."–"Yes, venerable sir," those bhikkhus replied. The Blessed One said this:

"And what, bhikkhus, is dependent origination? With ignorance as condition, volitional formations come to be; with volitional formations as condition, consciousness; with consciousness as condition, name-and-form; with name-and-form as condition, the six sense bases; with the six sense bases as condition, contact; with contact as condition, feeling; with feeling as condition, craving; with craving as condition, clinging; with clinging as condition, existence; with existence as condition, birth; with birth as condition, aging-and-death, sorrow, lamentation, pain, displeasure, and despair come to be. Such is the origin of this whole mass of suffering. This, bhikkhus, is called dependent origination.

"But with the remainderless fading away and cessation of ignorance comes cessation of volitional formations; with the cessation of volitional formations, cessation of consciousness; with the cessation of consciousness, cessation of name-and-form; with the cessation of name-and-form, cessation of the six sense bases; with the cessation of the six sense bases, cessation of contact; with the cessation of contact, cessation of feeling; with the cessation of feeling, cessation of craving; with the cessation of craving, cessation of clinging; with the cessation of clinging, cessation of existence; with the cessation of existence, cessation of birth; with the cessation of birth, aging-and-death, sorrow, lamentation, pain, displeasure, and despair cease. Such is the cessation of this whole mass of suffering."

This is what the Blessed One said. Elated, those bhikkhus delighted in the Blessed One's statement.