# An Enactive Approach to Value Alignment in Artificial Intelligence: A Matter of Relevance

Michael Cannon
Faculty of IE/IS
Philosophy and Ethics Group
Eindhoven University of Technology
Eindhoven, the Netherlands
m.a.cannon@tue.nl

## 1. Introduction

The "Value Alignment Problem" is the challenge of how to align the values of artificial intelligence with human values, whatever they may be, such that AI does not pose a risk to the existence of humans. A fundamental feature of how the problem is currently understood is that AI systems do not take the same things to be relevant as humans, whether turning humans into paperclips in order to "make more paperclips" or eradicating the human race to "solve climate change". Conceived this way, the problem is "how do we make relevant to AI what is relevant to humans?" Existing approaches do not explicitly address this feature. In this paper I make a proposal oriented toward solving the alignment problem understood in this way.

I begin by establishing some basic context, introducing the notion of the "existential risk" (xrisk) of AI, and noting the way in which value alignment is a connected with the problem of xrisk. I then offer a brief survey of several of the major existing approaches to value alignment, before proposing another approach. The proposal is philosophically rooted in the Enactive paradigm of contemporary cognitive science. I make two arguments to claim that, in order to align the values of AI with humans, we must make relevant for AI what is relevant to humans. The conclusion of this approach to the alignment problem is to make AI ontologically similar to humans.

Having made this claim, I note what Enactivism has to say about "making AI ontologically similar to humans". I suggest that the "4Es" of Enactivist cognition - embodiment, embeddedness, extendedness, and enactivity – each and all are conditions of cognition which define how certain things show up as relevant for humans. The implication is that to the extent that we design AI to be embodied, embedded, extended, and enactive, it will be aligned with human values because we will be more ontologically similar, and thus share a sensibility for what is relevant. The alignment is one in which the focus is not aligned solutions, but an aligned problem-space.

To conclude, I offer a summary comparison of this proposal with existing approaches, noting two things. Firstly, existing approaches and the approach proposed in this paper seem to have opposite ideas about how this relevance is defined. Secondly, existing approaches can be understood as solving for "low bandwidth alignment", whilst the proposed approach explores the conditions for "high bandwidth alignment". With this comparison made, an interesting trade-off presents itself, between the differential advantages of AI, and an increasingly value aligned AI. I will finish with a consideration of this tradeoff as it invites us to zoom out and reconsider what those values are with which we hope to align AI.

## 2. Value Alignment as a Solution to AI Xrisk: “You can’t fetch the coffee if you’re dead”

In this section I present the problem which contextualises value alignment - the existential risk of AI. Value alignment is a particular way of conceiving of the problem of AI existential risk. It is part of the growing field of work on “AI Safety”, the matter of building AI in a way that it does not harm humanity.

Its most developed and contemporary discussion is Brian Christian’s *The Alignment Problem: How can Machines Learn Human Values,* which engages with the question heavily in the context of state-of-the-art machine learning developments (2021). Important earlier discussions include Gabriel’s (2020) paper from *Deep Mind* and Nate Soares’ “The Value Learning Problem” (2016) from *The Machine Intelligence Research Institute*, both of which are more philosophical in their discussion.

Soares’ paper refers to the problem as the “value *learning* problem”, but it seems accepted that there are lots of different names, more and less colloquial or formal, for the kind of problem which improving AI seems to present.

The particular framing and conception of the problem as understood by “value alignment” is as follows. Following the work of Bostrom (2012, 2014) and Omohundro (2007) and, more recently, Russell (2019) there is an established concern that AI could become more intelligent than humans, and that this would be an existential risk to humanity, a risk to the existence of humanity. The thinking goes that, given even an innocuous goal, AI will recognise what is instrumentally necessary and rational to do in order to better achieve that goal, and this may bring it into conflict with humans. Russell captures this point in his quip that “you can’t fetch the coffee if you’re dead”:

> “If a machine pursuing an incorrect objective sounds bad enough, there’s worse. The solution suggested by Alan Turing – turning off the power at strategic moments – may not be available, for a very simple reason: *you can’t fetch the coffee if you’re dead.*
>
> …Suppose a machine has the objective of fetching the coffee. If it is sufficiently intelligent, it will certainly understand that it will fail in its objective if it is switched off before completing its mission. *Thus, the objective of fetching the coffee creates, as a necessary subgoal, the objective of disabling the off-switch.* The same is true for curing cancer or calculating the digits of pi. There’s really not a lot you can do once you’re dead, so we can expect AI systems to act pre-emptively to preserve their own existence given more or less *any* definite objective…
>
> There is no need to build self-preservation in because it is an *instrumental goal* – a goal that is a useful subgoal of almost any original objective. *Any entity that has a definitive objective will act as if it also has instrumental goals***.**” (Russell 2019: 140-141) (italics my own)

Behaving and developing itself rationally according to these instrumental goals is likely to bring it into a rivalrous dynamic with humans, and therein lies the risk to humans. Between Bostrom and Omohundro's work, there is a fairly comprehensive set of instrumental goals, including, alongside self-preservation, things like resource-acquisition and self-improvement, a uniquely enigmatic and vital variable in the xrisk story. Self-improvement is simultaneous with the other instrumental goals meaning that all of AI's capacities are improving. The self-improvement is recursive too, meaning AI is improving its very capacity for self-improvement as it improves. The risk of coming into rivalry with an AI which is potentially orders of magnitude more intelligent than us is what Russell calls the "Gorilla Problem": "…the problem of whether humans can maintain their supremacy and autonomy in a world that includes machines which substantially greater intelligence." (Russell 2019: 132)

Where value alignment comes in is the sense that the problem is that AI does not have our values and, if it did, it would realise that instrumental extinction of humans is not a valuable or *relevant* way to "solve climate change", say. Ensuring that AI has values aligned with humans, whatever that looks like ontologically, is thus a way of navigating the difference between "solving the problem" and "solving it in the relevant way".

> "We might call this the *King Midas problem*: Midas, a legendary king in ancient Greek mythology, got exactly what he asked for – namely, that everything he touched should turn to gold. Too late, he discovered that this included his food, his drink, and his family members, and he died in misery and starvation. The same theme is ubiquitous in human mythology. [Norbert] Wiener cites Goethe's tale of the sorcerer's apprentice, who instructs the broom to fetch water – but doesn't say how much water and doesn't know how to make the broom stop."
>
> A technical way of this is that we may suffer from a failure of *value alignment* – we may, perhaps inadvertently, imbue machines with objectives that are imperfectly aligned with our own." (Russell 2019: 137) (italics in original)

Nate Soares puts it this way in his paper: "[t]he Sorcerer's Apprentice problem arises when systems' programmed goals do not contain information about all *relevant* dimensions along which observations can vary." (Soares 2016: 2) (italics my own) This amounts to saying that nothing in the specification "solve climate change" specifies a particular way to do it. Indeed, an important incentive in building AI is the hope that it *will* supersede us, seeing ways of solving problems we do not and cannot, and so there is incentive not to anchor AI to human conceptions of possible solutions. This is to say, we do not know all the relevant dimensions of the problem space and, in this regard, "xrisk of AI" is a self-awareness of the naivety of building a Prometheus to bring down fire for us, not knowing how it will change everything.

The consequence is, again, that difference between "solving the problem" and "solving the relevant way", the way which is aligned with our values, whatever they may be. If "systems' programmed goals do not contain information about all relevant dimensions along which observations can vary", then we get situations where "get rid of humans" is a frustratingly rational way of solving climate change.

This is then the context of the Value Alignment problem. Value alignment is a strategic approach which imagines that if we can align the values of AI with our own, then AI will not just solve the problem, but ideally solve it in the relevant way, thereby mitigating the existential risk of AI.

I will now briefly present some of the existing approaches to value alignment and suggest that what they share is a sense that the way to go about things is to find the appropriate specification of the problem.

## 3. Some existing approaches to Value Alignment

In this section I briefly survey existing approaches to value alignment and suggest that they might be categorised together by their shared position on the nature of the problem, namely, a matter of defining the function or "goal" in the appropriate way.

As I noted in the previous section, value alignment is particular approach to building Safe AI, AI that does not harm humanity. Most, but not all, existing approaches are formal and technical in kind (i.e. mostly computer science) (Amodei et al. 2016, Yampolskiy 2016). They include "top down", "bottom up", and "hybrid", machine learning approaches. Yampolskiy in particular (2016) advocates a computer science of "Safety Engineering" in which we design "provably Safe" computer programs. The *Machine Intelligence Research Institute* also has their own approach to things, conducting research in Decision Theory, aiming to identify formal mathematical descriptions of the problem which ensure that AI makes the relevant decision at each stage of the execution of program. Whilst AI Safety is perhaps mostly understood as a field of technical research, it involves a complex of orientations on the problem.

Alongside the more technical work, there is also the field of Machine Ethics, rooted in philosophy and moral philosophy, which begins by treating AI as an autonomous moral agent, a subject in its own right rather than as an object (Anderson and Anderson 2007, 2011). Accounting for a spectrum of agential capacities that AI has (Moor 2006**)**, the questions of machine ethics considers how such agents, artificial or otherwise, ought to behave, as ethical agents (Wallach and Asaro 2017). This work has received a fair bit of critique as an approach to AI Safety, notably from Yampolskiy ("Machine Ethics is not the right approach") as well as a field of research in its own right (Brundage 2014). With regard to AI Safety and Value Alignment, the critique is usually includes on the one hand a claim that it is not adequately informed by, and responsive to, the concrete, technical work in AI – a criticism of naivety. On the other, for those who do engage it, the concern is philosophical, that machine ethics makes substantial assumptions about the nature of agency of autonomous systems of various kinds, from robots to algorithms, and so its philosophical inferences and conclusions are too speculative to legitimately inform strategy in the context of the existential stakes and risk of AI.

Russell's (2019) work on the "control problem" is an important integration of both the more technical aspects and the more philosophical aspects of the challenge. Being a computer scientist himself, his work on the question of how to remain in control of AI as it becomes increasingly intelligent and autonomous is sensitive to the most contemporary methods and developments in technical work on AI safety and value alignment, whilst also thinking philosophically about some of the major concepts, like "intelligence", (which are often otherwise assumed in a way that shows up in conditional statements like "if the AI is sufficiently intelligent, then it will realise…" ). I focus mostly on Russell's work from hereon, in part because of the way he tends to both technical and philosophical "sides" of the alignment problem, and largely because it seems to be quite representative of the character of existing approaches to alignment. These "sides" come out neatly when Russell says "[h]aving a reasonable definition of intelligence is the first ingredient in creating intelligent machines. The second ingredient is a machine in which that definition can be realized." (ibid: 32).

Recall from the first (indulgently) long Russell quote in the previous section:

> "… the objective of fetching the coffee creates, as a necessary subgoal, the objective of disabling the off-switch". (Russell 2019: 140-141)

It both explicitly establishes a recognition that the given goal specifies relevant sub-goals, and more generally and implicitly, that what is relevant to consider can be specified, derived, or inferred if the the goal or objective is appropriately articulated. This is how I suggest charactersising existing approaches to value alignment, that they conceive of the alignment problem as a matter of defining the appropriate goal or function such that what is relevant naturally falls and follows from it.

An important nuance of Russell's approach in particular is to define the goal or function in a way that includes uncertainty about what the goal is. In his proposal for "Beneficial Machines", he notes:

> The standard model underlying a good deal of 20th century technology relies on machinery that optimizes a fixed, *exogenously specified objective*. As we have seen, this model is fundamentally flawed. It works only if the objective is guaranteed to be complete and correct, or if the machinery can easily be reset. Neither condition will hold as AI becomes increasingly powerful.
>
> If the exogenously supplied objective can be wrong, then it makes no sense for the machine to act as if it is always correct. Hence my proposal for beneficial machines: machines whose actions can be expected to achieve our objectives. Because these objectives are in us, and not in them, the machines will need to learn more about what we really want from observations of the choices we make and how we make them. Machines designed in this way will defer to humans; they will ask permission; they will act cautiously when guidance is unclear; and they will allow themselves to be switched off. (Russell 2019: 247) (italics my own)

In this way, Russell's approach is "meta". He suggests we should not supply a (1st order) fixed objective, but supply a means of discerning what the objective itself is, that, if you will, the objective should be the discernment of what the objective is. His proposed means is to make the machines initially uncertain about what we do want them to do. He identifies three principles to this end (2019: 173):

1. The machine's only objective is to maximize the realisation of human preferences.
2. The machine is initially uncertain about what those preferences are.

3. The ultimate source of information about human preferences is human behaviour.

He goes on to explain that:

> ""Putting in values" is, of course, exactly the mistake I am saying we should avoid, because getting the values (or preferences) exactly right is so difficult and getting them wrong is potentially catastrophic. I am proposing instead that machines learn to predict better, for each person, which life that person would prefer, all the while being aware that the predictions are highly uncertain and incomplete." (Russell 2019; 178)

Russell has understood that 1st order specifications of the objective can lead to more problems, in the style of King Midas and so has stepped up a level of abstraction to a 2nd order approach. However, this 2nd order approach is still an expression of what I am suggesting is characteristic of existing approaches - that "if we find the appropriate (2nd order) articulation of the objective, then we can reach value alignment".

My concern ultimately is not one of critique. My point is that existing approaches, even one with nuance such as Russell's and those which echo it, as I understand for example MIRI to be doing, are of a kind, and that that kind is focused on defining the objective or "problem" to be solved , one way or another.

## 4. The role of "problem-solving" conceptions of intelligence in existing approaches to value alignment

I have mostly just stipulated that existing approaches have the character I suggested and have only offered a quick view of Russell's work to substantiate that. To support this, I will now briefly note how intelligence is understood in existing approaches to value alignment. Once the operational conception intelligence is clear, it becomes the clearer why existing approaches to value alignment treat it in terms of a problem or objective to be solved.

Citing the usual disclaimer about what is possible in the space of the paper, it is not feasible to make the case that each approach I noted is of the kind that I am suggesting. The operative conception of intelligence makes for a robust proxy for this though because it very naturally leads to approaching the value alignment problem as matter of defining the objective in the right way. This is to say, insofar as these approaches subscribe to a particular conception of intelligence, it follows from that particular conception that they would emphasise "defining the problem to solve in the appropriate way" to be the solution to the value alignment problem. Establishing that existing approaches subscribe to this conception of intelligence from which existing approaches follow is easier than showing that they subscribe to the category I am stipulating as "existing approaches". That is the first reason I am bringing up the matter of intelligence. It also makes for a nice segue into an introduction of the enactive-inspired approach, as the operative understanding of intelligence there is something different indeed.

As quoted earlier, Russell notes that "a reasonable definition of intelligence is the first ingredient in creating intelligent machines." (2019: 32) The prevailing understanding of intelligence in question is one which conceives of it effectively as a generalised "problem-solving" capacity: "It is this general problem-solving ability that we have in mind when we talk about "artificial general intelligence" (AGI) or "smarter-than-human AI."

The main features of this conception of intelligence are a focus on "problem-solving", and some kind of generalisation of broad/cross-domain quantification of that capacity. There is a third feature of these conceptions which I want to emphasise, and that is what they *don't* have, or what (must) they assume. They assume the identity of the problem and agent. This is important because *defining the problem* is not included in the conception of intelligence. I discuss the significance of this later in the context of enactivism.

Legg and Hutter's synthesis of 70-some definitions of intelligence is a common citation:

> "Intelligence measures an agent's ability to achieve goals in a wide range of environments." (2007: 9)

Achieving goals and solving problems amount to the same thing. In his 2012 paper and in his 2014 book, *Superintelligence,* Bostrom's articulations express something similar:

> For our purposes, "intelligence" will be roughly taken to correspond to the capacity for instrumental reasoning … Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal. (Bostrom 2012: 73)
>
> "By "intelligence", we here mean something like skill at prediction, planning, and *meansend reasoning in general*. This sense of *instrumental cognitive efficaciousness* is most relevant when we are seeking to understand what the causal impact of a machine superintelligence might be." (Bostrom 2014: 107) (emphasis mine)

Defining the problem is not "most relevant" in his definition of intelligence. Instead, the identity of the problem is assumed, and intelligence resembles an instrumental optimisation to that end.

Russell spends more time circling around the evolving understanding of intelligence in AI, chronicling the history of computers and AI through several paradigms and identifying in the current "Modern AI" paradigm a conception of AI as "rational agents" (2019: 42, Russell and Norvig 2010). He charts a trajectory through an Aristotelian conception of rationality, game theory, Bayesian reasoning under uncertainty, and comes to a conception that, too, is focused on problem-solving: "[g]iven a purpose defined by a utility or reward function, the machine aims to produce behaviour that maximizes its expected utility or sum of rewards, averaged over the possible outcomes weighted by their probabilities." (Russell 2019: 54) Note that it too assumes the purpose as "given", even if it does come to include 2nd order matters like uncertainty about what that purpose is in its conception of intelligence.

Generalised problem-solving is an intuitive conception of intelligence. My purpose here is not to critique it. See (Müller and Cannon 2022) for a paper-length discussion of this definition of intelligence as it pertains to the xrisk of AI. Rather, I want to bring out how a conception of intelligence which speaks of "solving" a problem,  without including or speaking to a capacity to "define" what the problem is to begin with, will lead to AI that struggles to realise what the relevant problem to solve is. Further, it also makes sense then that how to ensure that AI *does* solve the relevant problem, is to refine our articulation of the "problem".

Put slightly differently, if we do not include in our conception of intelligence the capacity to *define* what the ostensible problem is that we are to solve, and instead conceive of intelligence only in terms of problem-solving, it makes sense that a singularly problem-solving kind of artificial intelligence might be capable of solving the problem, but not in the relevant way. It has not properly understood what the problem really is. Given this behaviour, it also makes sense that we should then think that the way to solve any misalignment is to define the problem for AI.

Russell's Uncertainty approach offers "provably safe AI" insofar as AI won't unscrupulously act to optimise an objective it doesn't fully understand, but it doesn't guarantee that AI will necessarily come to any greater understanding. This is a point I will address in the last section of this paper.

Articulated as an argument, the claim I'm making looks something like this:

1. A conception of intelligence which assumes the identity of the problem to be solved leads to an understanding of the value alignment problem as a matter of defining the objective in the appropriate way.
2. Existing approaches to value alignment operate with this conception of intelligence.

C. Existing approaches to value alignment will be lead to an understanding of the problem as a matter of defining the objective in the appropriate way.

Again, I brought up the matter of intelligence in an attempt to offer something for my straight-up stipulation that existing approaches can all be put in the category of "define the problem the appropriate way". I acknowledge that with the second premise I am now effectively just stipulating of these approaches something else instead, namely, that they subscribe to a shared conception of intelligence as generalised "problem-solving". I am not leveraging the argument as a belief-forcing proof so much as explanatory aid. Though I do suspect that proponents of existing approaches would openly adhere to this model of intelligence, it is not in fact my concern. Identifying a connection between the operative conception of intelligence and a consequent perspective on value alignment, whether we take the connection to be logically necessary or not, reveals passage into a novel way of thinking about things. We know we will always struggle to account in the particular terms of a theory, for what that theory must assume to get off the ground.

I will now turn to a presentation of an approach to value alignment inspired by enactive cognitive science. The distinguishing feature for my purposes is that it engages the matter of how and why the previously assumed problem shows up as it does, with the parametres that it does. Enactivism suggests that there is a necessarily entangled relationship between subjects and objects, organisms and their "problems", a "dynamic co-emergence" of both. A sensibility for what is relevant becomes an inherent part of the conception of mind and intelligence. Consequently, the relevant approach to value alignment between humans and AI is to design AI to be a similar kind of subject to humans such that it perceives as we do.

## 5. Value Alignment informed by Enactive Cognitive Science: a matter of Relevance

The main attraction of thinking about value alignment from the perspective of enactive cognitive science is that is well-suited to speaking of relevance. It comes with the enactive model of cognition. (The enactive paradigm does not often speak of intelligence, so I will use the word "cognition", acknowledging that they are not identical terms, and gratefully asking the reader to abide.) In order to understand the enactive approach to value alignment, there are four key concepts of the enactive view to understand: the life-mind continuity thesis, autopoiesis, sense-making and dynamic co-emergence. These concepts articulate of condition of being in which a sensitivity to relevance is inherent to an agent. They are far from the entire enactive story, but suffice for my purposes and are already different in this way to existing approaches[1]. With that in place, I then present two arguments to get to the claim of this paper – for the greatest level of value alignment, we ought to design AI to be as ontologically similar to humans as possible. This conclusion is not ultimately the most interesting thing about an enactivist approach though. It is after all just another solution to a problem whose definition I am assuming, which amounts to an exact re-enactment of existing approaches. Perhaps the more interesting thing about the enactive approach is thus not the answers it offers, but the questions it demands.

The first concept, the "life-mind continuity thesis" is a vital, biology-inspired premise of the enactive view (Thompson 2007: chapter six). The idea is that where there is life, there is mind, and vice versa. "According to this thesis, life and mind share a set of basic organisational properties, and the organisational properties distinctive of mind are an enriched version of those fundamental to life." (ibid: 128) There are stronger and weaker versions of the thesis. In his textbook introduction to the philosophy of cognitive science, Andy Clark says of the life mind continuity thesis:

> "the thesis of strong continuity would be true if, for example, the basic concepts needed to understand the organisation of life turned out to be self-organisation, collective dynamics, circular causal processes, autopoiesis, etc., and if *those very same concepts and constructs* turned out to be central to a proper scientific understanding of mind" (Clark 2001: 118)

[1] A well and truly fantastic resource for a (60 page) paper-length, in-detail presentation of the enactive perspective, is Tom Froese and Tom Ziemke's (2009) paper, "Enactive Artificial Intelligence: Investigation the systemic organization of life and mind". It even presents the enactive paradigm in terms of design principles for building "enactive AI".

In terms that are relevant to this paper, the significance of the life-mind continuity thesis is that what is relevant for a mind is defined by its condition as a living organism. This leads to the second concept. Autopoiesis (self-production) can be understood as the dynamic condition of being a metabolically precarious organism, involving internal norms of self-organisation, regulation, and production. It is the self-production of an identity (Froese and Ziemke 2009: 25).

The self-production of a cell is the commonly used example to illustrate the idea, "...a cell produces its own components, which in turn produce it, in an ongoing circular process" (Thompson 2007: 98). In that way that self-organisation troubles the linearity of causation, a cell creates the conditions for itself to continue producing conditions for itself – "organisational closure" (Froese and Ziemke 2009: 26) or, in the context of complexity science "constraint closure" (Kauffman 2019). It is telling that Thompson (2007), Froese and Ziemke (2009), and Kaufmann (1993, 2019) all cite the same passage in Kant's *Critique of Judgement* when speaking about self-organisation, in which Kant differentiates machines and organisms, noting that an organism is both "cause and effect of itself", but a machine is not. Autopoiesis is this "constitutive autonomy" which generates "intrinsic teleology", internal norms of self-regulation and production, in terms of which discernment is made (Froese and Ziemke 2009: 28).

Autopoietic entities like a cell are thermodynamically open systems with semi-permeable boundaries which means they continuously must regulate the passage of matter and energy across that membrane boundary. Moreover, autopoietic entities must maintain their integrity – they must maintain themselves – in the face of the perturbations which this passage and metabolism of matter and energy effect. Autopoietic systems do this endogenously. This is the "constitutive autonomy" mentioned above. It is the significant difference between machines which, strictly speaking, are also open systems – machines are exogenously maintained, and do not autonomously manage their condition as open systems[2].

Now, this metabolising of matter and energy demands a sensibility for what is relevant to the continued existence of the individual, keeping certain substances within the cell to support the autopoietic process, and excreting others as waste products.

> "An organism must subordinate every change it undergoes to the maintenance of its identity and regulate itself and its interactions according to the internal norms of its activity. Life is thus a self-affirming process that brings forth or enacts its own identity and makes sense of the world from the perspective of that identity. The organism's "concern", its "natural purpose," is to keep on going, to continue living, to affirm and reaffirm itself in the face of imminent not-being." (Thompson 152-153)

There is always a sensibility for what is relevant because an autopoietic entity is always discerning in terms of what is relevant to its survival. This discernment necessarily involves a definition of the "problem", and always in a way that is relevant to the integrity and survival of the autopoietic self.

This discernment is the third concept, sense-making. Thompson defines it as "…the exercise of skillful know-how in situated and embodied action" (2007: 13), which is notably similar to Dreyfus' Heiddegarian "coping" (Dreyfus 2007) but not all that illuminating for anyone not already versed in

[2] Thanks to reviewer Carina Prunkl for inviting this nuance.

these languages. The point is that meaning, significance, and relevance are all inherent to sense-making. Sense-making is “meaning-generation”. In this way, normativity is already part of it.

One way of making sense of sense-making is as *adaptive discernment*, discernment of those differences, in both self and environment, relevant to the integrity and survival of the autopoietic self. In the condition of a precarious, semi-permeable identity, it is adaptive to be able to discern what is relevant, internally and externally, to one’s survival.

An important subtlety at this point is that this discernment is first and foremost of that which is relevant and therefore very much mind-dependent. It is of minor concern whether that which is perceived might be Real in some mind-independent way. In a recent textbook on enactivism Stewart notes in this regard that enactivism is “radically constructivist” (Stewart 2014: 27).

This is not a position necessarily shared by all enactivists, but it does facilitate intuitions for the last and perhaps trickiest concept, “dynamic co-emergence of autonomous selfhood and world” (Thompson 2007: 60). Consider the statement that “sense-making is the enaction of a meaningful world *for* the autonomous agent” (Froese and Ziemke 2009: 28) or, earlier, “[a] cognitive system is a system whose organisation defines a domain of interactions in which it can act with relevance to the maintenance of itself…” (Maturana 1970: 13) or from Merleau-Ponty, whose phenomenology directly inspires enactivism, (quoted here by Thompson) “[t]he environment emerges from the world through the actualisation or the being of the organism – [granted that] an organism can exist only if it succeeds in finding in the world an adequate environment” (Thompson 2007: 59).

Statements like this can make it sound an organism creates an external world, which is not quite the intended meaning. Recall the saying that we do not perceive the world as it is, but as we are; we do and can only perceive that which is relevant to us. We cannot perceive that which is not relevant to us. To the extent that, and in the way in which we do perceive it, it makes a difference to us and is thus relevant to us. Sense and adaptation is made. Enacted. It is such an understanding of the relation between selfhood and world which leads enactivists to say that “being defines a domain of relevance”. It means that the domain of what is relevant for a being – aka “the world” - is defined by the nature of that being. Each organism perceives what is relevant and adaptive for it to perceive. That is its “world”.

With these four concepts as basis for understanding the enactive conception of cognition, it is now possible to make sense of an enactive approach to value alignment. I make two arguments to generate a sense for an enactive proposal for value alignment.

## Argument for 1st Claim

1. Value for an entity is a function of what is relevant for that entity.
2. What is relevant for an entity is a function of the kind of ontological being it is.

C. Value for an entity is a function of the kind of ontological being it is.

This is simply the idea that different things are relevant to different kinds of beings, including what they value. Put this way, it suggests that the Frame Problem (Dennett 1984, Shanahan 1997, 2016) asks a very similar question to the alignment problem because both are an inquiry into relevance.

Shanahan (2016) notes that there are narrower and more general senses of the frame problem, but that an appropriately general understanding of it is the question of how to specify what is relevant in a given context. This is all but identical with the value alignment problem à la King Midas and the difference between "solving the problem" and "solving the relevant problem". A frame defines what is and is not in "the picture", what is and is not relevant.

In the enactive paradigm, an organism enacts its own frame. The "constitutive autonomy" of autopoiesis define the frame of reference in which both self and world emerge as meaningful. "Being defines a domain of relevance" can now be understood in terms of the way in which beings define – enact - a frame of reference (their "world", that which is relevant to them).

A comparison with the Frame Problem, borne of this exploration of enaction, invites an interesting rearticulation of the alignment problem as *"how do we make relevant to AI what is relevant to humans?"* I propose the following argument in response:

### Argument for 2nd Claim

1. Value for an entity is a function of the kind of ontological being it is. (1st Claim)
2. If beings are ontologically similar beings, they will have shared - i.e. aligned – values.

C. If AI and humans are ontologically similar beings, AI and humans will have shared – i.e. aligned – values.

Another way of putting the second claim is in terms of the Frame problem. If beings have the same or shared frame of reference, they will have similar values. If autopoietic identity is that which defines the frame, then that is how we can establish shared frame of reference.

Following from this claim, the suggested approach to alignment therefore is to design AI to be as ontologically similar as possible to humans - "ontological alignment".

Before continuing, it must be noted that nothing about this guarantees complete value alignment. As I discuss in the next section in which I explore what "ontological alignment" might look like, being similar beings far from guarantees alignment. Relative to other creatures, humans are very similar to one another, and yet still, even without AI, ourthe nuclear weapons we have built express the existential threat we pose to ourselves. Ontological alignment does not guarantee value alignment.

## 6. Designing the Frame for "AI to be as ontologically similar as possible to humans"

What does "design AI to be as ontologically similar as possible to humans" look like? I tentatively define it as the question of how to make relevant to AI what is relevant to humans? The Froese and Ziemke (2009) paper, again, is an excellent resource, more or less engaging this question directly. My offerings here will be a simplification that hopefully affords some grasp for intuitions. I suggest that in order to make relevant to AI what is relevant to humans, we have to build AI such that the same constraints define their “domains of relevance” and frames of reference as do ours. This is what I mean by “ontological alignment”. If we are similar kinds of beings, (ontologically aligned), our problem-spaces or “domains of relevance” will be similar. To this extent, the relevant *definition* of a problem, like climate change and inequality, may be aligned.

What are the relevant features of our being such that, if shared by AI, we might be “ontologically aligned”. I point at the so-called “4Es” of enactve cognition as a preliminary working suggestion for which constraints might be relevant.

The “4E’s” of cognition are embodiment, embeddedness, extendedness, and enaction (Newen et al. 2018). The important thing here is how these features define our frame, defining degrees of freedom through which meaning can be discerned and choice made. Sometimes these features are engaged in terms of how they “contribute” to cognition, which is to say, how they contribute to “solving problems”. Whilst this may be fruitful, it misses the matter of how the constraints of our condition generate a dynamic co-emergence of self and world.

Being *Embodied* in a physical form (seems to) localize and situate our awareness in 4-dimensional spacetime, grounding a “here”, and situating us in a context. Such a situation means being *Embedded* in ecological and evolutionary environments of all sorts, including social, mimetic, and technological environments as well as physical. Being *Extended* beyond our bodies (eg: via our smartphones and tools) means that our difference-making capacities and sense-making are not bound by the limits of our bodies, that our “worlds” are not bound by what our bodies can do and touch. Finally, being *Enactive* means being a precarious, *living*, metabolic, adaptive complex system.

All these conditions of being generate a “domain of relevance”. As such, they suggest themselves as design principles, at least, for building AI for which the same thing is relevant as for humans. Which is to say, value-aligned AI. Again, sharing these four conditions of being by no means guarantees a “Safe” outcome if applied to building AI. It does not guarantee that solutions to any problem will be identical, any more than they are when different humans look at a problem, but it does begin to address the King Midas matter of at least understanding how to specify what the *relevant* problem might be.

## 7. Summary Comparison: Low and High Bandwidth Alignment

In this section I want to zoom out and make a comparison of the earlier existing approaches I pointed to with the enactive approach I have just been suggesting. There are two points of comparison. I present these before asking whether there is a trade-off somewhere, and then finally concluding.

The first point of comparison between existing approaches and an enactive approach is what each assumes about how what is relevant is defined. In current “you can’t fetch the coffee if you’re dead” approaches, *what is relevant is defined by the problem,* whilst for the enactivism-informed approach, *what is relevant is defined by the being*. The arrow points in opposite directions across the subject-object split (Stewart 2014).

That said, the approaches are compatible. This leads me to the second point of comparison.

The approaches might be imagined along a spectrum from “low bandwidth” alignment, in which case AI *reasons* the same way humans do, but is still a machine, to “high bandwidth” alignment, in which case AI and humans are ontologically aligned by insofar as we are ontologically similar. An obvious challenge for a “high bandwidth” approach is that it makes for a much harder challenge than just figuring out the decision theory. I have represented this spectrum comparison in the figure below.

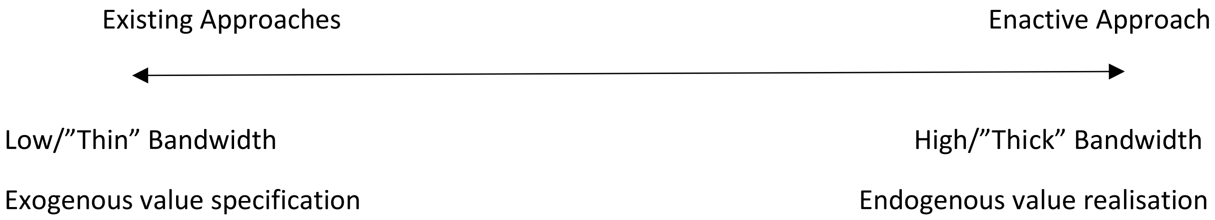

*Fig. 1 Bandwidth spectrum of approaches to AI Alignment*

### 7.1. A trade-off between Alignment and Differential Advantage?

One subtle challenge that this comparison suggests is a sense that maybe there is a trade-off between value-aligned AI which is Safe, and the differential advantage of AI as a tireless machine, which is not. This trade-off may be specific to an enactive approach to alignment in which Safe AI becomes ontologically more and more similar to humans and thus loses differential advantage. It nonetheless demands of us a more conscious inquiry into what we want from AI, which is no bad thing.

## 8. Conclusion

I began the paper with an introduction to the value alignment problem, situating it in the context of the existential risk of AI. I characterised the matter of value alignment as the difference between "solving the problem" and "solving the relevant problem", the frustrations of which are detailed in myths and legends like that of King Midas. I then discussed some existing approaches to alignment, categorising them together insofar as they imagine that alignment will happen if we can define the objective for AI in the appropriate way. I looked at Russell's uncertainty proposal in a bit more detail, but found that his move from 1st order specification of the objective to 2nd order means he is ultimately still doing the same thing. Following this discussion, I suggested that this way of thinking falls out of a conception of intelligence which sees it primarily as "problem-solving" and has to assume the definition of the problem in question.

From there I gave a quick presentation of some key concepts of the enactive paradigm, in which discerning and defining what is relevant is inherent to the model of mind. I suggested that the relevance of relevance for the enactive approach reveals an interesting connection to (and relevance of) the Frame Problem, which in turn affords a potent re-conception of the problem of alignment as "how do we make relevant for AI what is relevant for humans?" With this in mind, I then gave two sequential arguments to make the claim that in order to do so, we should to make AI as ontologically similar to humans as possible.

I finished up by comparing existing approaches with this proposal noting, most importantly here, that existing approaches offer “low bandwidth” alignment whilst the enactive approach offers “high bandwidth” alignment. This is turn seems to present something of a trade-off between value-aligned AI and AI with great differential advantage, but it might be particular to the enactive approach and not relevant for existing approaches.

Ultimately, if the enactive approach to value alignment offers anything of interest, it is perhaps not in any answer it offers, but it the kind of questions it demands of us to ask in this context, and the reconceptualisations of the “problem” it affords. That is, enacting an enactive approach means making sense of just what the problem might be.

## References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. *ArXiv:1606.06565 [Cs]*. http://arxiv.org/abs/1606.06565

Anderson, Michael and Susan Leigh Anderson. (2007). “Machine Ethics: Creating an Ethical Intelligent Agent”, *AI Magazine*, 28(4): 15–26.

- (eds.), (2011), *Machine Ethics*, Cambridge: Cambridge University Press. doi:10.1017/CBO9780511978036

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 355–372. https://doi.org/10.1080/0952813X.2014.895108

Christian, B. (2021). *The Alignment Problem: How Can Machines Learn Human Values?* New York: Atlantic Books.

Clark, A. (2001). *Mindware. An Introduction to the Philosophy of Cognitive Science*. Oxford University Press.

Dennett, D. (1984). Cognitive wheels: The frame problem of AI. In C. Hookway (Ed.), *Minds, Machines and Evolution* (pp. 129–151). Cambridge University Press.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology* 20 (2):247 – 268.

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, *173*(3–4), 466–500. https://doi.org/10.1016/j.artint.2008.12.001

Gabriel, I. (2020). Artificial Intelligence, Values and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford: Oxford University Press.

- (2019). *A World Beyond Physics: The Emergence and Evolution of Life*. Oxford: Oxford University Press.

Maturana, Humberto R. (1970). Biology of Cognition. *Autopoiesis and Cognition: The Realization of the Living*, *43,* (Boston Studies in the Philosohy of Science), 2–58.

Müller, V. C., & Cannon, M. (2022). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio*, *35*(1), 25–36. https://doi.org/10.1111/rati.12320

Moor, J. H. (2006). “The Nature, Importance, and Difficulty of Machine Ethics”, *IEEE Intelligent Systems*, 21(4): 18–21. doi:10.1109/MIS.2006.80

Newen, A., Bruin, L. D., & Gallagher, S. (Eds.). (2018). *The Oxford Handbook of 4E Cognition*. Oxford, UK: Oxford University Press.

Omohundro, S. M. (2007). *The Nature of Self-Improving Artificial Intelligence*.

Russell, S. (2019). Human Compatible: AI and the Problem of Control. UK: Allen Lane.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach: International Edition* (3rd ed.). Pearson.

Shanahan, M. (2016). The Frame Problem. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016). https://plato.stanford.edu/archives/spr2016/entries/frame-problem/

- (1997). *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. MIT Press.

Soares, N. (2016). The Value Learning Problem. *Machine Intelligence Research Institute*, 7.

Stewart, J., Gapenne, O. & di Paolo, E. (Eds.). (2014). *Enaction: Toward a New Paradigm for Cognitive Science*. Cambridge, MA: MIT Press.

Stewart, J. (2014). “Foundational Issues in Enaction as a Paradigm for Cognitive Science: From the Origin of Life to Consciousness and Writing” in *Enaction: Toward a New Paradigm for Cognitive Science*. Stewart, J. and Gapenne, O., and Di Paolo, E. A. (eds). Cambridge, MA: MIT Press.

Thompson, E. (2007). *Mind in Life*. Cambridge, MA: Harvard University Press.

Wallach, W. and Peter M. A.  (eds.), (2017). *Machine Ethics and Robot Ethics*. London: Routledge.

Yampolskiy, R. V. (2016). *Artificial Superintelligence: A Futuristic Approach*. Boca Raton: FL: CRC Press.

“*Why AI Safety?* “ (n.d.). Machine Intelligence Research Institute. Retrieved 11 April 2022, from https://intelligence.org/why-ai-safety/