Intelligent Behaviour

Dimitri Coelho Mollo

dimitri.mollo@umu.se

Umeå University, Department of Historical, Philosophical and Religious Studies

& Cluster Science of Intelligence, Berlin

Preprint

Accepted in Erkenntnis – *Please cite only the published version*

Abstract

The notion of intelligence is relevant to several fields of research, including cognitive and comparative psychology, neuroscience, artificial intelligence, and philosophy, among others. However, there is little agreement within and across these fields on how to characterise and explain intelligence. I put forward a behavioural, operational characterisation of intelligence that can play an integrative role in the sciences of intelligence, as well as preserve the distinctive explanatory value of the notion, setting it apart from the related concepts of cognition and rationality. Finally, I examine a popular hypothesis about the underpinnings of intelligence: the capacity to manipulate internal representations of the environment. I argue that the hypothesis needs refinement, and that so refined, it applies only to some forms of intelligence.

1 Introduction

The notion of intelligence is notoriously difficult to characterise precisely. Part of the reason for this is that the concept of intelligence lives many different lives, some pretheoretical, some within a variety of scientific disciplines. In psychology, the perhaps dominant approach has been to come up with ways to measure human intelligence, mostly by means of various types of tests, and then try to explain individual variation in the scores obtained by appeal to differences in underlying capacities (Nęcka & Orzechowski 2004, Sternberg & Pretz 2004). Intelligence tests often include tasks that concern linguistic and mathematical capacities, excluding thereby from consideration nonhuman forms of intelligence that do not employ external symbolic systems. Furthermore, a consensus about what it is that such tests measure is lacking.

Human intelligence is also the benchmark in Artificial Intelligence, which relies on rarely articulated common-sense understandings of intelligence — an approach inherited, perhaps, from the influential but problematic Turing Test (Turing 1950). In comparative psychology and ethology, different degrees of intelligence are ascribed to different animals in light of the capacities they display, although clear criteria are typically lacking, or are overly anthropocentric (see Hurley & Nudds 2006*b*). Moreover, in animal cognition research, as in philosophy, intelligence is commonly conflated with the related notions of cognition and rationality, with a few notable exceptions (Hurley & Nudds 2006*a*, Fridland 2015).

This rather confusing situation, in which intelligence is characterised differently within and across scientific fields, let alone in everyday usage, calls for a close examination of the notion of intelligence. One way to go is to argue that, for the reasons just mentioned, intelligence is too vague and diverse a notion to really

find a home in science. We should rather eliminate or replace it with something better, clearer and more tractable. I cannot help but feel the lure of such a position, especially once one is reminded that intelligence research is a relatively old field, going at least a century back, and still struggling to understand what it is supposed to be about. Eliminativism's allure notwithstanding, I will argue that the notion of intelligence is scientifically and philosophically fruitful, and that there are promising ways to circumscribe it so as to make it explanatorily and theoretically useful.

I propose five basic *desiderata* that a notion of intelligence should meet if it is to play valuable epistemic and pragmatic roles in the sciences that make use of the notion. In light of these *desiderata*, I provide a substantive, practice-oriented view of intelligence that can be shared among the many sciences of intelligence, paving the way for cross-disciplinary interaction and integration. Interaction and integration which, I take, are crucial to make progress in shedding light on the nature, roles, and varieties of intelligence in the biological and artificial realms.

Here is how I will proceed in what follows. In section §2 I present and defend five *desiderata* for a satisfying characterisation of intelligence, and I briefly delve into the limited, but substantial agreement on some core features of intelligence that can be gleaned from existing work on intelligence. I put forward my operational, behavioural characterisation of intelligence in section §3. I expand this defence in section §4, proposing a distinction between intelligence, cognition, and rationality. Finally, in section §5 I argue that the influential hypothesis that intelligent behaviour is explained by computational processing of internal representations needs further refinement, and applies only to some forms of intelligence.

2 Delineating Intelligence

Some *desiderata*, and some agreement

In order to try and shed light on the distinctive features of intelligence, as well as on its explanatory role in the relevant sciences, it is important to look at the characterisations of intelligence offered by scientists working in those fields. This does not mean that we need to take their word as writ in stone — a doomed stance to take, given the many extant disagreements. Looking for a fruitful, cross-disciplinary characterisation of intelligence — or anything more ambitious than that — is not a matter of accommodating intuitions or pretheoretic assumptions, not even experts' intuitions and assumptions. It is rather a matter of identifying a notion of intelligence that can play a distinctive theoretical and explanatory role — both as *explanans* and as *explanandum* — in the relevant sciences. Examining the characterisations suggested by practitioners of those sciences offers a promising way to understand what roles the notion of intelligence is supposed to play in them. It remains however an open possibility that closer analysis may reveal that no such theoretically satisfying notion is available, or that there are a plurality of different notions in different fields, with no unification possible. Intuitions in one direction or another may thus be flouted, and bear very little, if any, epistemic weight.

Before we take a closer look at extant characterisations of intelligence, it is worthwhile to get a better grip on what we should be looking for, if our aim is to identify a basic, shared, scientifically-fruitful understanding of the notion that can make space for integration between the different relevant fields of research. These aims suggest the following *desiderata* for a notion of intelligence:

Species-neutrality An adequate notion of intelligence should not be character-

ised in such a way as to *a priori* privilege one species over others. This can also be called the non-anthropocentrism *desideratum*, since the tendency has historically been that of giving pride of place to what *prima facie* only humans can do as the criterion for intelligence, such as language comprehension, playing chess, and doing predicate logic. Anthropocentrism about intelligence is problematic on several counts. First, it constrains without justification the realm of what is intelligent to the specific cognitive and motor abilities, needs, interests, and ecological embeddedness of a single biological species. Second, and most centrally for my purposes, anthropocentrism denies that there can be a cross-disciplinary notion of intelligence covering also the behaviours and abilities of nonhuman systems — biological and artificial — that ethologists, animal cognition researchers, comparative psychologists, and AI researchers are interested in. If humans are taken as the standard, we may end up failing to properly study and understand the varieties of intelligent behaviour, if any, that follow from cognitive and bodily make-ups that are radically different from the human one, or to do so only by means of comparisons to a standard that has little to do with the needs, goals and ecologies of nonhuman systems.

Origin-neutrality Analogously to **species-neutrality**, an adequate notion of intelligence should not be characterised in such a way as to *a priori* privilege systems with one type of origin. This is crucial to avoid excluding non-biological systems, such as potential future artificial systems, from counting as intelligent, or downplaying their intelligence. Moreover, a notion of intelligence that applies exclusively to biological systems, or that takes them as standard, risks failing to make space for alternative forms of intelligence

that may be proprietary to non-biological systems, given their possibly very different abilities, goals and aetiology (i.e. natural selection vs. design). Such a biology-centred notion of intelligence would ignore or downplay potentially rewarding avenues of research into forms of artificial intelligence possibly very unlike biological ones.

Multiple realisability An adequate notion of intelligence should not be characterised in such a way as to *a priori* privilege one kind of implementing mechanism. It is possible, and given current knowledge very plausible, that intelligence is multiply realised: that is, brought forth by a diversity of different underlying mechanisms. The case is clearer when reading 'mechanism' structurally: we have good reason to think that different physical structures, biological and artificial, individual and collective, can underlie intelligence. This is likely also true on a more abstract, functional reading of mechanism: there may be different organisations of internal (or collective) functions, however implemented, that can give rise to intelligence. While there is a chance that this may prove false, for instance if it turns out that having a certain functional organisation or implementing a certain set of computations is necessary and sufficient for intelligence, it certainly should not be assumed from the get-go. Thus **multiple realisability** should remain as a *desideratum*, which may though be rejected in light of future empirical and theoretical work.

Epistemic distinctiveness An adequate notion of intelligence should be such that it plays a distinctive theoretical and explanatory role in the relevant sciences. That means that the notion must add something of epistemic value to scientific theories and explanations. It should not be merely a ter-

minological variant of other useful notions, such as cognition and rationality. It should thus be useful for categorising a distinctive set of phenomena of interest, which call for distinctive theories and explanations. This does not mean that we need commit to the idea that intelligence corresponds to a natural kind in the world, or to any neatly delimited type of phenomenon. In order to satisfy this *desideratum*, a notion of intelligence must organise scientific theorisation and experimentation in fruitful ways, open up promising paths of research, and offer useful tools for explaining, measuring, and modelling the phenomena of interest. None of this entails realism about intelligence as a natural kind, let alone any type of essentialism. It does not entail their falsity either.

Responsiveness to scientific practice An adequate notion of intelligence should be such that it is not overly revisionary, that is to say, it should respond to and refine the use already made of the notion in the relevant sciences. As mentioned above, this does not mean just accepting the extant characterisations of intelligence offered by practitioners, as they may flout some or all of the above desiderata, and their simple combination, given the existing disagreements, would lead to incoherence. This desideratum does require, however, that extant characterisations and the use they are put to should inform an adequate notion of intelligence. They are good points of departure for an adequate notion of intelligence, but not the point of arrival.

In sum, a notion of intelligence that can be a plausible candidate for allowing cross-disciplinary integration of the relevant sciences, and for being scientifically fruitful should, as a minimum, be species- and origin-neutral, allow multiple realisability, be epistemically distinctive, and responsive to scientific practice. In the next

section, I will propose a characterisation of intelligence that meets these *desiderata*. First, however, I will take the cue from the last *desideratum*, and look at extant characterisations of intelligence, so as to use them as stepping stones to my own proposal.

A variety of definitions and informal characterisations of intelligence have been provided throughout the hundred years or so of dedicated scientific work on intelligence across different fields, several of which collected in Legg & Hutter (2007*a*)[1]. I will leave it to the reader to peruse those definitions. For my purposes, it suffices to point out that while it is often noted that a shared, accepted characterisation of intelligence is lacking within and across the various fields that study biological and artificial intelligence, it is less often recognised that there is a considerable degree of agreement across the board on several features taken to be central to intelligence. A large number of proposed definitions of intelligence make explicit or implicit reference to at least one, and often several of the following factors:

- *Generality*, or displaying appropriate behaviour in a wide range of different tasks and circumstances;
- *Flexibility*, or adjusting behaviour in light of changing and/or uncertain circumstances;
- *Goal-directedness*, or forming and pursuing goals appropriate to the circumstances;
- *Adaptivity* (adaptive learning), or informing current and future behaviour in light of past interactions with the world[2].

[1] See also Legg & Hutter (2007*b*), Fridland (2015), Lake et al. (2016), Hurley & Nudds (2006*a*), Shevlin et al. (2019), Marcus (2020).

[2] Other factors that are less often mentioned include: metacognitive abilities (Sternberg & Pretz 2004, Nęcka & Orzechowski 2004), neural efficiency (Schulz 2010), and environmental and

So as to respect the **responsiveness to scientific practice** *desideratum*, these largely agreed-upon core features of intelligence should inform attempts to come up with a scientifically-fruitful cross-disciplinary characterisation of intelligence. The stage is now set for my positive proposal.

3 An operational, behavioural characterisation of intelligence

Before I go on, it is important to keep in mind that my aims are partially but crucially different from those that spurred several of the extant definitions of intelligence. I am not looking for a definition in any strict sense of the word: no set of necessary and sufficient conditions that capture the 'essence' of intelligence, nor conditions the fulfilment of which is criterial for something to count as intelligent. In part this is due to the fact that I am deeply sceptical of the existence of strict definitions and criteria for most interesting, not merely stipulated scientific concepts, especially in the special sciences. Most importantly, the phenomena that the notion of intelligence tries to capture appear to be too fuzzy, diverse, and variable for them to be non-trivially accommodated by a definitional or criterial straitjacket.

My ambitions are much more modest, but also, I take, more fruitful. I aim at an operational, minimal characterisation of intelligence that fulfils the five *desiderata* set out above, each of which has independent justification, as we have seen. The

symbolic scaffolding (Dennett 1996, Deacon 1997). For a similar breakdown of the core features of intelligence, see Fridland (2015). She identifies adaptive learning in particular as crucial, the other three features being explainable in terms of it. I will remain neutral on this question, although I believe that none of those four features are conceptually or explanatorily more central than the others. I take them rather to form an interdefinable cluster of concepts, with goal-directedness being the most primitive and least specific. At any rate, none of what follows hinges on this.

desiderata on species- and origin-neutrality and multiple realisability, in particular, forbid any characterisation that privileges one or a few species, biological over artificial systems (or vice-versa), or specific kinds of mechanistic or functional realisations. They invite a characterisation that is purely behavioural, and thus neutral on all those respects[3]. A behavioural characterisation has moreover the extra, pragmatic bonus of relative simplicity of detection: behaviour, after all, is relatively easy to observe and test. It is not as easy to categorise, however, and thus a behavioural characterisation of intelligence should help to individuate some sorts of behaviour as the ones that are indicative of intelligence — providing, as it were, a behavioural symptomatology of intelligence.

Paying heed to the considerations above about the core features of intelligence typically mentioned in existing research, I venture the following operational, purely behavioural characterisation of intelligence:

> An intelligent system S is a system that manifests behaviours[4] that are often enough
>
> - *general*, i.e. that are appropriate in a variety of different circumstances;
> - *flexible*, i.e. that change appropriately in light of changed, novel, or uncertain circumstances;
> - *goal-directed*, i.e. that are appropriate in light of goals S plausibly possesses;

[3]Hernandez-Orallo (2017) also focuses, for similar reasons, on behavioural factors as central to characterising intelligence, especially when it comes to psychometrics. However, he employs a broader notion of behaviour, which for him includes personality traits and cognitive processing, than the one operative here, which makes reference exclusively to overt behaviour.

[4]It is also possible to cash out this characterisation in terms of dispositions to behave, at least under Vetter's (2014) possibility conception of dispositions, which individuates dispositions purely in terms of their manifestations, rather than of stimulus-manifestation pairs, as per the traditional view.

- *adaptive*, i.e. that change appropriately in light of previous interactions with the world.

Being purely behavioural, and embodying the four core features of intelligence that play a central role in most existing definitions of intelligence — in their behavioural translations, as it were — the foregoing characterisation clearly fulfils four out of the five *desiderata* it was meant to satisfy, namely **species-neutrality**, **origin-neutrality**, **multiple realisability**, and **responsiveness to scientific practice**. Manifesting behaviours that have those four features fits well with the apparent target of the notion of intelligence in cognitive science, artificial intelligence, ethology and comparative psychology (and possibly also folk intuitions).

The characterisation above does not involve any direct or indirect requirement that intelligent systems be biological or artificial, that they belong to specific species, or that they possess a specific kind of mechanistic or functional organisation — thereby leaving open the possibility that intelligence is multiply realised both structurally and functionally. Moreover, it suggests relatively straightforward 'tests' for the presence of intelligence: systems that manifest often enough behaviours with those characteristics in their habitats, under experimental conditions, or perhaps in simulated environments, are to be operationally characterised as intelligent.

The proposed characterisation has two further features worth mentioning. It is somewhat fuzzy, insofar as what counts as 'often enough', and what counts as similar or different circumstances cannot be determined precisely, and certainly not *a priori*. Such judgements depend on case-by-case assessment of the epistemic value and reasonableness of ascriptions of intelligence. Systems that rarely display behaviours with the relevant features are plausibly to be placed outside the fuzzy

boundaries of the domain of intelligence. Clearly to be excluded are one- or few-shot cases in which those features of behaviour are manifested, but for which there are better explanations in terms of coincidence or luck. Similarly, systems that behave appropriately only in a certain kind of task — e.g. recognising objects in different scenes — do not display the form of generality in behaviour required by the operational characterisation. However, borderline and unclear cases are to be expected, given the fuzzy nature of the proposed characterisation, and of the phenomena it tries to capture.

The foregoing characterisation is fuzzy along another dimension, insofar as the symptoms of intelligence it individuates are themselves a matter of degree. In particular, patterns of behaviour may be more or less general, and more or less flexible. This fact indicates two paths worth treading.

A conservative option is to hold that there is a threshold, albeit vague, under which the degrees of generality and flexibility in behaviour become such as not to count as symptoms of intelligence any more. Above this threshold we find different degrees of intelligent behaviour. The threshold should be set in light of pragmatic and epistemic considerations; in terms, that is, of the pragmatic and epistemic fruitfulness, or lack thereof, of ascribing intelligence, and different degrees of intelligence, to systems in different locations in this graded space.

An alternative option is to reject the idea of a threshold, and embrace a liberal view of intelligence, according to which most or all biological and artificial systems are intelligent, albeit most to a relatively lower degree, and few to a higher degree. This, I believe, might be the strategy preferred by those philosophers and scientists that have recently been exploring the idea that intelligence can be found in relatively simple systems, such as bacteria, fungi, and plants.

I will argue for the first, conservative view in the next section. After all, it will

not have escaped the reader's attention that I have been silent so far on one of the *desiderata*, namely **epistemic distinctiveness**. Respecting this *desideratum*, I will show, puts quite some strain on the liberal option.

Before going further, however, an important aside is in order: it is worthwhile to set aside a classical objection to behaviour-based definitions of intelligence, put forward by Block (1981). Block famously criticised behaviour-based definitions of intelligence by imagining what came to be called Blockheads: systems comprising giant look-up tables or tree structures that include all possible sensible behaviours given any possible input, in addition to a brute force string or tree search computational procedure to find the appropriate output for each input. Block argues that it is clear that such systems are not intelligent, even though their behaviour would be indistinguishable from a human's, concluding that behaviour-based definitions of intelligence are thereby misguided.

While Blockheads may be conceivable, they are nomologically impossible, as Block himself (mostly) recognises. It is also debatable whether intuitions about such outlandish cases should be given much weight. At any rate, the sort of behavioural characterisation I am suggesting here is neither definitional nor criterial: it is operational, motivated mostly by epistemic and pragmatic considerations. Whether Blockheads are conceivable, nomologically possible or else is thereby irrelevant to the foregoing proposal. Operational characterisations should be expected to fail in some cases, especially in highly contrived or atypical ones, such as Blockheads[5]. This does not jeopardise the epistemic and pragmatic value of the characterisation, provided that it proves to be fruitful in its application to actual

[5]Similarly for cases in which intelligent systems are permanently prevented from behaving (e.g. due to locked-in syndrome). Ascription of intelligence to such systems must rely on additional considerations, having to do, for instance, with whether the system is a token of a type that typically displays intelligent behaviour, whether we have grounds to believe that the dispositions to behave are still present but cannot be manifested, etc.

biological and artificial systems, as well as to realistic potential future systems.

4 The distinctive role of intelligence

The notion of intelligence is typically, and justifiably, used in conjunction with the notions of cognition and rationality. The boundaries between the notions are often unclear. There is little agreement within and across fields about whether these notions capture different phenomena at all, intelligence being sometimes conflated with cognition, sometimes with rationality. The lack of terminological agreement is in itself unhelpful, as it may hinder cross-disciplinary exchange in an area of research that is by its very nature interdisciplinary. Regimenting terminology is thereby valuable and important. But I believe that there is more to the issue than mere terminology.

These three notions, I hold, play distinctive explanatory roles in our sciences of cognition and behaviour. For when properly constrained, they capture specific sets of phenomena that pose different questions, call for different methods of investigation, and play different explanatory roles. In sum, in order to respect the **epistemic distinctiveness** *desideratum*, a characterisation of intelligence should be non-trivial, and such that it sets intelligence apart from the notions of cognition and rationality.

Let us start with the notion of cognition. As Fridland (2015) points out, in the cognitive sciences cognition is applied quite broadly, encompassing most or all of the internal processes in the nervous systems of organisms that help to inform behaviour, including thereby sensation, perception, various forms of learning, memory, among many others. In mainstream cognitive science, these processes are typically taken to consist of computations being performed over

internal representations (Godfrey-Smith 1996, Dennett 1996, Penn et al. 2008, Schulz 2010, Marcus 2020).

Internal representations are states of a system that carry information about or stand in for external states — such as the presence of food or predators in the environment — and are used by the system or its subsystems in virtue of that, so as to guide appropriate behaviour. The conditions of adequacy of representations — that is, the conditions that determine whether a representation is true or false, satisfied or else — are representational contents. The content of a representation is what the representation is about, or, more intuitively, what it means. Often an additional requirement is added: only representational contents that depend on the history, interactions, and workings of the system itself are taken to be relevant to characterising cognition — representational contents that depend on the intentions and purposes of other systems are excluded (Adams & Aizawa 2008 phrase this requirement in terms of non-derived content, see also Rowlands 2009)[6].

Alternative accounts deny that cognition need involve computations over representations. (Varela et al. 1991, Hurley 1998, Noe 2004, Thompson 2007, Lyon 2017)[7]. Such approaches see cognition as involving, minimally, processes of self-maintenance and self-organisation that are the bread and butter of life: maintaining the internal organisation and boundaries of the organism, preserving its unity in its dealings with its environment, and partly thanks to such dealings. Cognition, on this picture, is widespread in the biological realm, from bacteria to

[6]In consequence, this additional requirement is incompatible with pragmatist and fictionalist theories of representation (Egan 2014, Sprevak 2013, Coelho Mollo 2020), while it is arguably compatible with views that are realist about representations but not about representational vehicles (Dennett 1981, 1991), as well as with forms of semirealism about representation (Coelho Mollo forthcoming).

[7]Some of these philosophers also defend the more radical claim that the body and sometimes the environment are themselves part of cognitive states and processes. These more radical approaches typically adopt the label '4E cognition', which stands for embedded, embodied, enacted, and extended cognition.

plants to mammals — albeit coming in many gradations, from less (bacteria) to more complex (humans). On this view, there are forms of cognition that may not involve computations and representations, while it remains a possibility that some forms of cognition do.

It is well beyond the scope of this paper to settle this debate, which came to be known as the quest for 'the mark of the cognitive' (Adams & Aizawa 2008, Adams 2010, Rowlands 2009, Lyon 2017, Sims 2021). What should be rejected, I take, are overly demanding, narrower characterisations of cognition, which have been common especially in philosophy of mind. According to such views, cognition comprises only processes involving conceptual thinking, reasoning, planning, and other capacities that humans typically excel at (Fodor 1975, Adams 2010). This has led to what Hurley (1998) has dubbed the 'sandwich picture' of the mind. Perception and action are the two loaves of bread, between which lies the cognitive filling: the central executive processes that truly constitute cognition. Cognition, on this understanding, largely corresponds to intelligence (Fridland 2015).

As Sims (2021) argues, such narrow views of cognition fail to capture the practices of cognitive science. Indeed, the domain of phenomena that cognitive scientists investigate include capacities very unlike the ones delineated by such narrow views, such as non-flexible, automatic, and stereotypical behaviours (e.g. perceptual biases, behavioural routines); innate, non-learned capacities and behaviours (e.g. innate learning biases and conceptual structures[8]); domain-specific capacities and behaviours (e.g. cognitive modules and behavioural habits[9]); non-conceptual and subpersonal states and processes (Shea 2018); associative and reinforcement learning; in addition to the more complex capacities involved in reasoning, lan-

[8]See e.g. Mandler (2004), Carey (2009).
[9]See e.g. Carruthers (2006), Fridland (2019).

guage, and planning that narrow views privilege (Lyon 2017).

A broader understanding of cognition is thereby to be preferred. Importantly for my purposes, a broad notion of cognition, motivated by the foregoing considerations, allows us to carve up smaller domains of cognitive phenomena that have distinctive features. Intelligence, I hold, is one such domain.

Therefore, cognition is best understood broadly, capturing capacities that are common in phylogeny and ontogeny. I will here remain neutral on which specific view of the mark of the cognitive is to be preferred among the broad ones[10]. In either case, cognition is a process-based notion: it refers to capacities of biological and artificial systems that involve certain kinds of processes, be them computational processes over representations, or processes of self-maintenance and self-organisation. Cognition is thereby an undemanding notion, which might be applicable to plants, fungi, bacteria, should ongoing empirical research provide compelling evidence that they undergo the appropriate kinds of processes picked out by the preferred mark of the cognitive.

In consequence, the notion of cognition includes those systems to which a liberal view of intelligence may want to ascribe low levels of intelligence. This indicates that using the notion of intelligence in this liberal way has little epistemic value, insofar as claiming that such systems have low levels of intelligence does not add much of epistemic and pragmatic value over and beyond the claim that they are cognitive. For this reason, I prefer the more conservative view of intelligence mentioned in the previous section. If the proponent of the liberal view should introduce additional distinctions to capture the relevant quantitative and qualit-

[10]Even though I think that the representational view is more promising, insofar as views based on self-maintenance and self-organisation risk failing to accurately capture cognitive scientific practices by going to the other extreme in comparison to narrow views: they make cognitive science encompass an overly rich domain of phenomena, including most of biology.

ative differences between what, by their lights, would all be instances of intelligent phenomena, then the liberal view would amount to little else than a terminological variant of the conservative view.

Importantly, the conservative view does not exclude by *fiat* the possibility that plants, fungi, bacteria, and other similar systems might be intelligent. It just requires them to meet more demanding standards — manifesting general, flexible, goal-directed, adaptive behaviour often enough — than the ones to be met to count as cognitive. Whether such systems meet those standards or not is of course an empirical question.

Rationality, in contrast, captures a much narrower domain of phenomena. As Kacelnik (2006) points out, there is considerable variation over how to characterise rationality across different fields — all of them narrower than the characterisations of cognition (and intelligence). In philosophy and psychology, he argues, rationality is a process-based notion involving deliberation and reasoning in the formation of beliefs and belief-like states. In other fields, such as economics and evolutionary biology, rationality is, on the contrary, purely or mostly an outcome-based notion, focused on the extent to which systems' behaviours respect or approximate optimality constraints fixed by a normative model, typically involving reward or fitness maximisation (Kacelnik 2006). Even in its process-based characterisation, the normativity involved in judgements of optimality is central to the notion of rationality, for which metacognitive processes of error detection and confidence estimation may be required (Hurley & Nudds 2006*a*, Stanovich 2012). In addition, the normativity of rationality may come from many different sources, several of which tied to anthropocentric economic, social, and moral values (Pepperberg 2006).

My contention is that intelligence lies somewhere in between cognition and

rationality. Some intelligent systems are cognitive systems, and some intelligent systems — those that respect normative and/or optimality constraints and involve additional error-sensitive processes — are rational. I remain neutral on whether all intelligent systems are cognitive. While the claim seems plausible, a definite answer depends on the open question of the mark of the cognitive. Should we take processes of self-maintenance and self-organisation to be characteristic of the cognitive, there might be computational-representational systems that are general, flexible, adaptive, and goal-directed but that lack those processes — they would thereby be intelligent, but not cognitive. On the other hand, should we take the presence of computations over representations as central to cognition, then it becomes more likely that all intelligent systems are also cognitive, since the core features that characterise intelligence arguably require such processes (see section §5 below).

Rational systems, in their turn, may not be all located within the set of intelligent systems, at least if we accept pointillist rationality, that is, narrow domain-specific rationality: think of DeepBlue or AlphaGo and their rational choices of winning moves in chess and Go. In consequence, while there are some intelligent systems that are rational, there are (pointillist) rational systems that are not intelligent[11].

The relationships between cognition, intelligence, and rationality are illustrated in figure 1.

If these considerations are on the right track, intelligence turns out to have its own distinctive theoretical and explanatory identity, occupying a specific place in the conceptual repertoire of the relevant sciences. Its being more demanding than

[11] And not even cognitive, if such systems, as seems plausible, fail to compute over non-derived representations, or fail to feature processes of self-organisation and self-maintenance.

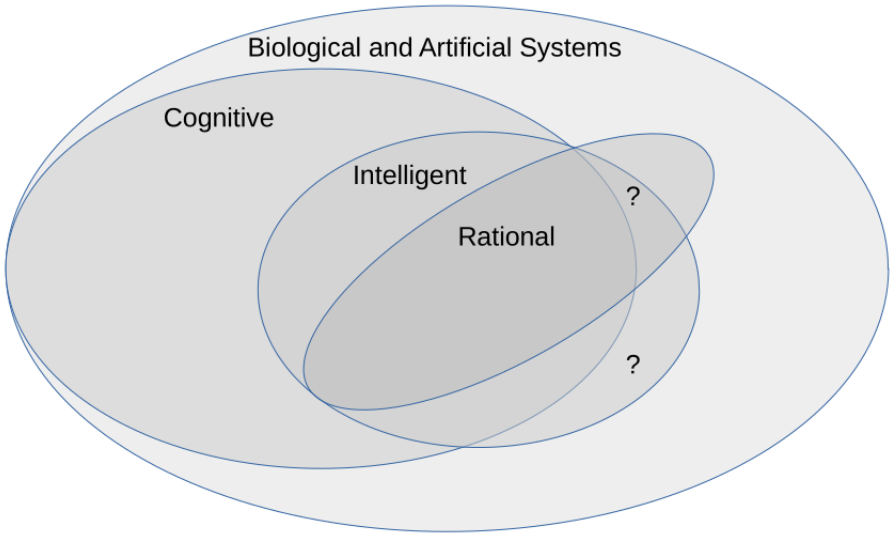

Figure 1: Relationships between cognitive, intelligent, and rational biological/artificial systems. Question marks indicate areas that may not be occupied.

the notion of cognition, furthermore, indicates that the characterisation of intelligence provided is not trivial. It seems thus that we have good grounds to claim that the foregoing behavioural characterisation of intelligence meets **epistemic distinctiveness** as well.

5 Looking under the hood

I have defended a behavioural, operational characterisation of intelligence that, I have argued, meets the five *desiderata* aimed at unifying and organising cross-disciplinary research on intelligence. An intelligent system, I have claimed, is one that manifests, often enough, behaviour that is general, flexible, goal-directed, and adaptive. Before closing, I would like to examine a popular hypothesis about the underpinnings of intelligence in mainstream cognitive science, according to which intelligent behaviour is produced by computational processes over internal

representations (Godfrey-Smith 1996, Lake et al. 2016, Garson 2019). I will try to show that this hypothesis, if it is to be fruitful and informative, needs to be further refined; and that so refined, it provides at best partial explanations of certain forms of intelligence.

Importantly, the question at hand is different from the one that occupied me in previous sections, namely that of providing an operational characterisation of intelligence that can be fruitful across the sciences interested in intelligent phenomena[12]. Here, instead, my aim is to evaluate a claim about how the behaviours that characterise intelligence may be realised in intelligent systems. In other words, while the characterisation of intelligence is purely behavioural, in order to satisfy the **multiple realisability** *desideratum*, once we determine that a system or a family of systems fulfils the characterisation, we can then ask by what means they are able to display intelligent behaviour. Such an explanation will typically involve specific kinds of processes, among which a prominent candidate is the computational processing of representations.

The idea that intelligence involves representations is very influential, being part and parcel of the mainstream commitment in cognitive science to representational accounts of cognition, whereby internal representations are appealed to in explaining cognition as a whole, as we have seen above. In light of the considerations in section §4, however, appealing to the use of representations *tout court* as distinctive to explanations of intelligence would be problematic.

If we accept the mainstream view that cognition is characterised by the performance of computational processing over (non-derived) representations, appealing to representation in explaining intelligence would not be distinctive of intelligence, since such an appeal is also called for in explaining cognition. If we endorse

[12]I am indebted to an anonymous reviewer for urging me to clarify this point.

alternative, non-representational views of cognition instead, the situation does not improve. While such views arguably put a lower bound to cognition than representational views, they are compatible with the claim that many paradigmatic cognitive phenomena involve representations, such as the ones listed in the previous section. Those phenomena, however, include non-flexible, non-general, non-learned behaviours. In consequence, once again, appeal to representations *tout court* would not add anything distinctive to an explanation of intelligence, since representations are already appealed to in explaining some forms of non-intelligent cognition.

A more promising suggestion regarding the relationship between representation and intelligence, I take, is to appeal to demanding kinds of representational processing in explaining certain kinds of intelligence. The modified hypothesis thus becomes the following: representational processing of certain kinds provides a plausible, partial explanation for certain kinds of intelligence, in particular at the high flexibility, high generality end of the spectrum. This does not preclude that there may be a plurality of different explanations that apply to different families of intelligent systems, and/or to different forms and degrees of intelligence[13]. What kind of representational processing might be the relevant one?

Let us go back to the four core features of behaviour that characterise intelligence, namely generality, flexibility, goal-directedness and adaptivity. Let us see which properties representational processing must have, such that it can help generate behaviour with high degrees of those features.

The relevant computational processing of representations must be such that it exploits information about the environment stored in representations, putting that information to use, if need be, in different circumstances (*generality*). In

[13]Thus the **multiple realisability** *desideratum* is not jeopardised.

addition, it must be such that the stored representations are modified in light of experience in such a way as to make future behaviour more appropriate and/or more likely to be successful (*adaptivity*). The relevant representational processing must also be such that it appropriately combines stored representations in novel ways, employing them to deal with new, changing and/or uncertain circumstances (*flexibility*). Finally, the relevant processes must be such that they can produce and use directive representations, that is to say, representations that represent goals to be achieved; or more weakly, that they help guide behaviour toward achieving the system's goals — regardless of whether the goals are explicitly represented (*goal-directedness*)[14][15].

In order to fulfil these requirements, the representations produced and used must be storable, reusable and combinable at least to some degree. One way in which representational systems can have these features to a high degree is by having or approximating compositional structure, which allows productivity and systematicity. That is to say, the representational system must produce and use representations that are approximately discrete, that are to some degree context-insensitive, and that can be combined following principles of composition, such that together they can produce an indefinite number of composite representations whose contents are a function of the contents of the component representations. Natural languages are paradigmatic examples of such compositional, productive representational systems, motivating Fodor (1975) to posit a Language of Thought (LOT), or Mentalese, a proposal closely related to the traditional symbolic ap-

[14]I thank an anonymous review for bringing this point to my attention.

[15]Another motivation often adduced for appealing to internal representations is that they increase the efficiency of information storage and decrease computational load. Representations allow systems to use compressed versions of the rich information gathered from the environment, selected in terms of its behavioural relevance, thus helping tackle what is known in Artificial Intelligence as the 'curse of dimensionality' (Lake et al. 2016, Poldrack 2020).

proach to Artificial Intelligence (Newell 1980).

I am not arguing for LOT specifically. It does fit the bill at hand, but it is likely not the only option do to so. After all, as I pointed out above, approximating compositional structure may well be enough. The representations may have fuzzy boundaries and overlapping parts, thus not being fully discrete. They might also be context-insensitive only relative to a more or less wide range of contexts, and their combination principles may take the shape of soft constraints, rather than rigid logical rules (Smolensky 1988). Sub-symbolic architectures, such as connectionist networks, can thereby possess the relevant representational features as well.

I have been qualifying my claims about computational processes over representations being explanatory of intelligence by saying that they are at best partial. Why so? There are two main reasons.

First, rich representational structures as the ones involved in quasi-compositional representational systems may not be needed for many sorts of moderately general and flexible intelligent behaviour. Map-like or tree-like representational structures, which fall short of the representational power and productivity of language-like representations, may suffice for flexible, adaptive, goal-directed behaviour with limited generality — for instance, behaviour that concerns only limited, albeit broad task domains, such as spatial and social navigation (Camp 2007, 2009, Boyle 2019). Forms of intelligent behaviour that involve even less generality and flexibility can plausibly be underpinned by yet simpler structures. In consequence, appeal to quasi-compositional representational processes contributes to plausible explanations of some sorts of intelligent behaviour only.

Second, and more generally, explanations that appeal to representational processing tend to abstract away from the roles played by external factors in bringing about and making possible intelligent behaviour. Representational structures can

be, and often are, augmented by external objects and practices, such as language, maps, social conventions, items and landmarks in the environment. Representational storage can be offloaded onto environmental arrangements and external media, including other living beings. Moreover, what counts as general, flexible and adaptive behaviour hinges on the bodily capacities of systems, on the ways in which they can or cannot intervene on their surroundings (Clark 1998, Barsalou 2008, Brooks 2018). Similarly, generality, flexibility and adaptivity can only be evaluated in light of the needs and goals of systems, as well as of the possibilities they have for satisfying those needs and fulfilling those goals given their representational and bodily capacities, and the structure and state of the environment. In this sense, appeal to representational processing only provides part of the explanation for most, if not all, instances of intelligent behaviour.

What I have offered in this section, therefore, is just a small part of the puzzle. A puzzle that, in order to be properly solved, will require overcoming deep-seated divisions in the sciences of intelligence, such as those that oppose representationalism and embodiment; symbolic, connectionist, and embodied AI; and human and nonhuman biological intelligence. I hope, at any rate, that the operational characterisation of intelligence that I defended in the previous sections can help transcend these unhelpful divisions, thus contributing to a more integrated, cross-disciplinary approach to intelligence research.

6 Concluding Remarks

I have argued in favour of two main claims. First, I have proposed and defended a purely behavioural characterisation of intelligence, aimed at providing a notion of intelligence that is scientifically fruitful, explanatorily distinctive, and that

provides a shared basis for cross-disciplinary interaction and integration across the sciences of intelligence. Second, I have suggested that producing and using representations that approximate compositional structure is one plausible candidate for a partial explanation of intelligence, especially for what regards behaviours that score high in generality and flexibility.

While most of the ideas presented here have in one form or another been circulating in the philosophical and scientific literature for some time, they typically lacked careful and detailed philosophical examination and development, and were often not concerned specifically with intelligence (exceptions include Hurley & Nudds 2006*a*, Fridland 2015, Hernandez-Orallo 2017). I tried to fill a bit of that gap in this paper. I have also attempted to point out the much work yet to be done in this relatively underexplored, but particularly important, area of philosophical research.

Acknowledgements

I am indebted to Michael Pauen, Alex Kacelnik, Margherita Arcangeli, two anonymous referees, as well as to audiences at the 2nd International Conference in Philosophy of Mind and Cognition (Rio de Janeiro, 2020), at the Workshop Intelligent Abilities (Berlin, 2021), at the CEPE/IACAP Joint 2021 Conference (Hamburg), and at the Morning Talks Series at the Science of Intelligence Cluster (Berlin, 2020) for valuable feedback on earlier versions of this material. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

References

Adams, F. (2010), 'Why we still need a mark of the cognitive', *Cognitive Systems Research* **11**(4), 324–331.

Adams, F. & Aizawa, K. (2008), *The bounds of cognition*, Wiley, Malden, MA.

Barsalou, L. W. (2008), 'Grounded cognition', *Annual Review of Psychology* **59**, 617–45.

Block, N. (1981), 'Psychologism and behaviorism', *The Philosophical Review* **90**(1), 5.

Boyle, A. (2019), 'Mapping the minds of others', *Review of Philosophy and Psychology* **10**(4), 747–767.

Brooks, R. A. (2018), Intelligence without reason, *in* L. Steels & R. Brooks, eds, 'The Artificial Life Route to Artificial Intelligence', Routledge.

Camp, E. (2007), 'Thinking with maps', *Philosophical Perspectives* **21**, 145–182.

Camp, E. (2009), A language of baboon thought, *in* R. W. Lurz, ed., 'The Philosophy of Animal Minds', Cambridge University Press.

Carey, S. (2009), *The Origin of Concepts*, Oxford University Press.

Carruthers, P. (2006), *The Architecture of the Mind*, Oxford University Press.

Clark, A. (1998), *Being There*, MIT Press.

Coelho Mollo, D. (2020), 'Content pragmatism defended', *Topoi* **39**, 103–113.

Coelho Mollo, D. (forthcoming), 'Deflationary Realism: Representation and idealisation in cognitive science', *Mind & Language* .

Deacon, T. (1997), *The symbolic species : the co-evolution of language and the brain*, W.W. Norton, New York.

Dennett, D. C. (1981), True believers: The intentional strategy and why it works, *in* A. F. Heath, ed., 'Scientific Explanation: Papers Based on Herbert Spencer

Lectures Given in the University of Oxford', Clarendon Press, pp. 150–167.

Dennett, D. C. (1991), 'Real patterns', *The Journal of Philosophy* **88**(1), 27–51.

Dennett, D. C. (1996), *Kinds of minds : toward an understanding of consciousness*, Basic Books, New York, NY.

Egan, F. (2014), 'How to think about mental content', *Philosophical Studies* **170**, 115–135.

Fodor, J. A. (1975), *The Language of Thought*, Harvard University Press.

Fridland, E. (2015), Learning our way to intelligence: Reflections on Dennett and appropriateness, *in* 'Content and Consciousness Revisited', Springer International Publishing, pp. 143–161.

Fridland, E. (2019), 'Longer, smaller, faster, stronger: On skills and intelligence', *Philosophical Psychology* **32**(5), 759–783.

Garson, J. (2019), 'Review of efficient cognition: the evolution of representational decision making, (armin w. schulz, MIT press, cambridge, MA, 2018)', *Biology & Philosophy* **34**(3).

Godfrey-Smith, P. (1996), *Complexity and the Function of Mind in Nature*, Cambridge University Press.

Hernandez-Orallo, J. (2017), *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*, Cambridge University Press.

Hurley, S. L. (1998), *Consciousness in Action*, Harvard University Press.

Hurley, S. & Nudds, M. (2006*a*), The questions of animal rationality: theory and evidence, *in* 'Rational Animals?', Oxford University Press.

Hurley, S. & Nudds, M., eds (2006*b*), *Rational Animals?*, Oxford University Press.

Kacelnik, A. (2006), Meanings of rationality, *in* 'Rational Animals?', Oxford University Press, pp. 87–106.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. (2016), 'Building

machines that learn and think like people', *Behavioral and Brain Sciences* **40**.

Legg, S. & Hutter, M. (2007*a*), A collection of definitions of intelligence, *in* B. Goertzel & P. Wang, eds, 'Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms', Vol. 157, IOS Press, pp. 17–24.

Legg, S. & Hutter, M. (2007*b*), 'Universal intelligence: A definition of machine intelligence', *Minds and Machines* **17**(4), 391–444.

Lyon, P. (2017), 'Environmental complexity, adaptability and bacterial cognition: Godfrey-smith's hypothesis under the microscope', *Biology & Philosophy* **32**(3), 443–465.

Mandler, J. M. (2004), *The Foundations of Mind*, Oxford University Press.

Marcus, G. (2020), 'The next decade in AI: Four steps towards robust artificial intelligence', *arXiv 2002.06177* .

Nęcka, E. & Orzechowski, J. (2004), Higher-order cognition and intelligence, *in* 'Cognition and Intelligence', Cambridge University Press, pp. 122–141.

Newell, A. (1980), 'Physical symbol systems', *Cognitive Science* **4**, 135–183.

Noe, A. (2004), *Action in perception*, MIT Press, Cambridge, Mass.

Penn, D. C., Holyoak, K. J. & Povinelli, D. J. (2008), 'Darwin's mistake: Explaining the discontinuity between human and nonhuman minds', *Behavioral and Brain Sciences* **31**(2), 109–130.

Pepperberg, I. M. (2006), Intelligence and rationality in parrots, *in* 'Rational Animals?', Oxford University Press, pp. 469–488.

Poldrack, R. A. (2020), 'The physics of representation', *Synthese* .

Rowlands, M. (2009), 'Extended cognition and the mark of the cognitive', *Philosophical Psychology* **22**(1), 1–19.

Schulz, A. W. (2010), 'The adaptive importance of cognitive efficiency: an alternative theory of why we have beliefs and desires', *Biology & Philosophy* **26**(1), 31–

50.

Shea, N. (2018), *Representation in Cognitive Science*, Oxford University Press.

Shevlin, H., Vold, K., Crosby, M. & Halina, M. (2019), 'The limits of machine intelligence', *EMBO reports* **20**(10).

Sims, M. (2021), 'A continuum of intentionality: linking the biogenic and anthropogenic approaches to cognition', *Biology & Philosophy* **36**(6).

Smolensky, P. (1988), 'On the proper treatment of connectionism', *Behavioral and Brain Sciences* **11**, 1–23.

Sprevak, M. (2013), 'Fictionalism about neural representations', *The Monist* **96**, 539–560.

Stanovich, K. E. (2012), On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning, *in* K. J. Holyoak & R. G. Morrison, eds, 'The Oxford Handbook of Thinking and Reasoning', Oxford University Press.

Sternberg, R. J. & Pretz, J. E. (2004), Unifying the field: Cognition and intelligence, *in* 'Cognition and Intelligence', Cambridge University Press, pp. 306–318.

Thompson, E. (2007), *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, Harvard University Press.

Turing, A. M. (1950), 'Computing machinery and intelligence', *Mind* **59**(236), 433–460.

Varela, F., Thompson, E. & Rosch, E. (1991), *The Embodied Mind: Cognitive Science and Human Experience*, The MIT Press.

Vetter, B. (2014), 'Dispositions without conditionals', *Mind* **123**(489), 129–156.