Yesterday's tomorrow today: Turing, Searle and the contested significance of Artificial Intelligence

Chp 5 pp. 82-137 in Al-Amoudi, I. and Morgan, J. editors (2018) *Realist Responses to Post-Human Society: Ex Machina* (Volume 1) London: Routledge

Jamie Morgan

Introduction

Artificial intelligence (AI) has become an increasingly important issue in and for societies. It has also become entangled with what is termed Transhumanism (TH). In this paper I explore the way AI is conceived and focused upon (with some comment on TH). Two conceptual focuses of AI have emerged and these in turn have roots in and are related to key works that have dominated how AI has been addressed through philosophy. The key works are Turing (1950) and Searle (1980).[1] I explore the arguments of these two. The intent is not a pejorative 'back to basics', but rather an exploration of sophisticated origins in order to identify how dividing lines and omissions can become in some ways ingrained and in others interstitial. Both involve problems of ontology and social ontology, which in turn creates problems for how we seek to shape the future. From both Turing and Searle a weak and strong focus in AI has developed and this has had a variety of further consequences. The consequences are themselves complicated and inter-connected and so cannot be simply stated or enumerated but arise cumulatively as the argument proceeds. In the final section I draw the whole argument together in terms of the social significance of actual technological changes occurring under the aegis of AI, and do so finally with reference to the concept of relational goods (Donati and Archer, 2015).

Two concepts of artificial intelligence (AI)

For AI a convenient place to start is with the much-publicised first report from the Stanford *One Hundred Year Study on Artificial Intelligence*. The project was launched in 2014 and as the title suggests is a long-range study of AI. Its remit is to explore both the state of and consequences of AI. The project brings together a designated multi-disciplinary panel of experts every five years to provide an update on progress. As an authoritative exercise it expresses some typical positions within AI research.[2] The first report notes there is a 'lack of a precise, universally accepted definition' of AI (Stone,

[1] In stating the problem is not new and in drawing attention to Turing and Searle as longstanding sources of key ideas I do not mean to suggest that the history of AI, computing, robotics, animatronics and automata begins with these two. Neologisms may be new but ideas far older. John McCarthy coined the term AI at a conference in 1956. There are 'robots' in the *Iliad* and actual animatronics of varying degrees of sophistication are scattered throughout history. See Adam Rutherford 'Rise of the Robots: The history of things to come' Radio 4 broadcast Monday 13th February 2017: http://www.bbc.co.uk/programmes/b08crvz3

[2] There are recognized limitations. The first report restricts itself to impacts for a typical American city based on 8 domains. It explicitly excludes military and security aspects of AI and assumes that the material presented will be relevant to different degrees on a global basis.

2016: p. 12). However, this is not quite right. There are two basic conceptual focuses for AI and some confusion about how they relate. How something is conceived is usually more complex in its formation than how it is defined. Furthermore, what the conceptual focus emphasises has consequences for how a concept operates.[3] What I mean by this will become clearer as we progress.

For our purposes, Searle introduces the relevant basic conceptual distinction for AI in his 'Minds, brains and programs,' (1980). His concern was the significance of AI for the study of the mind. For Searle, 'weak AI' concerns computers as a tool for the study of the human mind, whilst 'strong AI' assumes or asserts that a sufficiently complex computer is a mind. We will return to Searle later. The distinction, though still partly rooted in Searle's work, has generalised beyond a focus on philosophy of mind. Weak AI focuses primarily on AI as functions and strong AI as entities.

Weak artificial intelligence (AI^{w})

In contemporary usage the weak concept of artificial intelligence (AI^{w}) focuses on function and is on closer inspection semantically minimal and tautological. The Stanford report adopts Nils Nilsson's well-known definition and neatly encapsulates key aspects of AI^{w}:

> Artificial Intelligence is that activity devoted to making machines intelligent and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment [... According to this view] the difference between an arithmetic calculator and a human brain is not one of kind, but of scale, speed, degree of autonomy and generality [… There is a] intelligence spectrum [… and] the characterization of intelligence as a spectrum grants no special status to the human brain. But to date human intelligence has no match in the biological and artificial worlds for sheer versatility, with the abilities to reason, achieve goals, understand and generate language, perceive and respond to sensory inputs, prove mathematical theorems, play challenging games, synthesize and summarize information, create art and music and even write histories [… But for AI] matching any human ability is only a sufficient condition, not a necessary one. (Stone, 2016: p. 12)

It should be emphasised that AI^{w} is a default position, and often a placeholder. The Stanford report, like many others, attempts to leave open exactly what AI might be and become. New reports are to be commissioned by a Standing Committee for the Stanford project every 5 years. The first report was produced by a selected Study Panel, mainly comprised of experts in robotics, programming, data analysis, systems theory and planning, and economics (drawn from Microsoft, MIT, Harvard etc). The default to AI^{w} reflects the lack of agreement and coherency across these fields. This lack, awkward though it seems, is not empty for the purposes of how something is

[3] To be clear, as an exercise in analytical philosophy, one could formally distinguish definition, concept and conceptual focus. In what follows I do not elaborate significantly regarding distinctions. The main point I am making is that definition starts from a simple statement about intelligence as a kind of doing, and that the concept of AI observably emphasises either function or entity, and so in a practical sense of development of concerns is bound up with focus, which in turn has consequences.

conceived, since it invites reduction, minimalism and tautology to create a point of departure: *the troubling problem of defining AI has been addressed and so we can move on*. This tacitly introduces a 'be' into AI[w]. If one reads the Stanford report it claims AI will be what AI researchers do and explore. So, AI becomes an accidental identity: *AI is AI*. Concomitantly, since there is a great deal of activity and development occurring under the rubric AI, what it does becomes the convenient focus for what it is. Thus, AI[w] focuses on function, albeit increasingly complex function, and what can be tested and (mainly) observed. Here, function is primarily an expression of 'intelligence'. The main subsidiary distinction becomes one between specific 'intelligence', defined as the more-or-less efficient or appropriate doing of something, and general 'intelligence', defined as the replication of this functionality in multiple, different and new domains of application.[4]

Many AI researchers are aware that the distinction between specific and general intelligence is based on a blurring of what 'intelligence' itself is. For example, Legg and Hutter (2007) collate around 70 available definitions of intelligence, and subcategorise these into those deriving from psychologists and from AI researchers. Their intent is to synthesise a human independent concept of intelligence amenable to AI researchers. They identify 3 primary features: 1) intelligence is a property of an individual agent as it interacts with an environment, which 2) relates to its ability to succeed with reference to an objective, and 3) depends on agent adaptability to objectives and environments. 'Intelligence' can then be measured via 'achievement', subject to 1-3. Clearly, since the paper is developed in the context of AI concerns, this definition can be read as tending towards a general intelligence concept of AI. Still, it remains heavily weighted towards AI[w]. Moreover, the derived concept of intelligence has been based on a particular selective synthesis. It notes but puts aside a whole set of key aspects of the psychologists' definitions and broader conceptual statements. What seems *common* is selected, whilst what may be important but only highlighted by some is omitted. Several non-AI definitions highlight the composite and interpretive basis of intelligence, and the range of aspects of intelligence stated by psychologists but omitted in the final definition includes: to plan, to think abstractly, to show good sense, practical sense and initiative through judgement and associated activity, to learn facts and skills and appropriately apply them and to demonstrate awareness of the relevance of behaviour. These aspects are sieved from the synthesis, and so the concept of intelligence is pre-structured in Legg and Hut's synthesis in a way that encourages a focus on function. AI as a focus of *practical* concerns can focus on little else than function, and so seeking a *common* definition is either circular or

[4] There have been major developments in specific 'intelligence', in ways that address some of the specific areas of complexity in the Stanford report quoted list of human achievements. The EMI program has successfully imitated the work of Bach. In a more high profile case, the DeepMind project at Google is responsible for, AlphaGo, which is now able to defeat a human Go world champion using Monte Carlo simulation and tree search within a system based on multiple self-play. See neural networks material later section and also:
https://deepmind.com/research/publications/mastering-game-go-deep-neural-networks-tree-search/ In general, machine learning using large datasets has changed some aspects of how a computer plays games. However, this is still currently far from general intelligence, even in the AI[w] sense, since the AI so far are unable to simply turn their 'intelligence' to a different setting or game etc. They must begin anew each time. This may change of course, but this in itself would not satisfy characteristics for strong AI.

self-propagating. The intent to develop something human independent becomes simply isolated (rather than abstracted) from important characteristics that may situate intelligence. Function, is the common concern that occurs across *all* definitions and so becomes the central aspect *of* intelligence.

Concomitantly, the focus on function in AI is a pragmatic response to potential based on *one* set of key contemporary concerns. It takes the present and expresses a future based on an engineer or designer's frame of reference. Since AI is not currently conscious etc. it seems convenient to concentrate on what identified 'AI' can do and may do rather than what a 'intelligent entity' *is* and may *be*.[5] There is a tacit and sometimes acknowledged semantic slippage involved since AI researchers sometimes refer to AI and then 'true AI', with the latter referring to a fully realised entity. Yet AI^{w} still involves a version of 'be'. AI^{w} adopts, often inadvertently, an external and behaviouristic approach to how AI is defined and so how AI is conceived. This is most readily understood in terms of specific 'intelligence', but is significant also for general 'intelligence'. Difference of kind is put aside in favour of position along a spectrum and this has consequences: a calculator and a human function differently and are differently complex, but are as 'intelligences' not properly distinguished constitutively or qualitatively. This ambiguity of distinction is basic to AI^{w}. It is, as we shall see, rooted in Turing's approach, though arguably Turing also made strong AI claims that encourage a focus on function via functionalism as constitutive for an entity.

One should note that the problem of kind for AI^{w} is not clear-cut. Read sympathetically in context, the typical intent of an AI^{w} default is to defer discussion of what AI may be in a constitutive-qualitative sense, until such time as it becomes germane. This is conceptually problematic and potentially dangerous, even if one puts aside doomsday Terminator scenarios. The previous quote from the Stanford report states that matching human abilities is sufficient but not necessary (SbnN) as a benchmark for AI. Given that AI has already been subject to a spectrum that encompasses function referenced to a calculator, SbnN introduces a component that affects the coherence of this way of *defining* AI. One might now infer anything less able than a human in its flexibility, diversity and complexity (i.e. AI^{w} general intelligence) would be insufficient to be AI.[6] The authors seem to mean no more than that the future abilities of AI *may* exceed the comparable abilities of humans and that given abilities may not be *restricted* to those of humans. This is still about function and so says nothing directly about what is *sufficient* or *necessary* to the *constitution* of 'intelligence' or to the human, despite that the human is stated as the benchmark.

There is thus a basic ambiguity in AI^{w} regarding what it is that is different, what this difference derives from and whether emergence is a meaningful concept to apply to both the human and AI (for context see Stephan, 2006).[7] It is perhaps worth noting

[5] Clearly a programmer or designer must give thought to what something is in order to make it, but this is in order to make it do something and is different than extended rumination or reflexive focus on what something is as an entity in terms of matters of status, kind, comparative constitution, qualities and so forth.

[6] And yet may contain something necessary to AI, so there is the possibility of insufficient necessary as well as sufficient non-necessary constitution. The problem is logical-semantic and so infects claims of substance formulated within the schema.

[7] The implication is not that intelligence emerges from prior forms such as calculators along a spectrum (since this would be absurd in various ways) but rather that any particular entity on the spectrum may be described in emergent terms.

that Darwin ultimately positioned humans and animals in the same way, so this is not a procedure that is especially controversial when dealing with distinctions. The problem it creates is that it seems to inadvertently equate entities that are distant on the spectrum, since the spectrum implies no breakpoints or thresholds for difference, even if one might want to acknowledge that a threshold could be broad and blurred rather than narrow and neat.

In the case of AI, the curious corollary of this AI^{W} functional approach is that it decentres what it is to be human or, to be less prejudicially anthropocentric, to be equivalent in terms of possible *essential* characteristics (since these invoke issues of kind and may be relevant matters to address in terms of some animals, aliens etc.). This is important because it is built into the conceptual construct, and so cannot readily be rectified by subsequently acknowledging the importance of the human whilst analysis and argument proceeds on the basis of AI^{W}. A conceptual disjuncture is liable to linger. This is also problematic, since the future is a matter of how technology is designed, shaped, and used by, in and for humans within societies. As the authors of, for example, the Stanford report are aware, what matters is how technology is shaped, and how technology in turn will shape society. Few AI scholars set out to be simple technological determinists. However, the disjuncture and focus on function does encourage a kind of tacit weighting *towards* characteristics of determinism: AI is coming and we have to *cope* with it. Concomitantly, the disjuncture creates problematic beginnings for how we deliberate regarding human futures based on what it is to be human, which can affect in turn how humans flourish or suffer. These are quintessentially issues of ontology.

Strong artificial intelligence (AI^{S})

Whilst AI^{W} focuses on function, in contrast strong artificial intelligence (AI^{S}) takes a step back to consider what directs function. AI^{S} thus locates 'intelligence' within an expanding set of characteristics which may be associated with this direction: purpose, awareness, cognitive unity, consciousness, self-consciousness etc. AI^{S} is mainly concerned with the nature of entities. It focuses on the constitution that affects external expression, and so mediates and enables function. Moreover, function is not the only concern, being merely a subset of the consequences of the existence of an entity. There are two main subcategories of an AI^{S} conceptual focus. First, one subcategory focuses on the equivalence between human and 'AI', and thus on the validity of analogical claims. This locates what it means to be intelligent within philosophy of mind. *Inter alia*, when posed as a machine-mind (program) problematic it invites disputes regarding what function alone can reveal or allow one to infer regarding the nature of mind, organic or otherwise. Though concerned *by* AI the main concern is *with* the human and what AI does or does not tell us about the human. This focus follows from Searle's work. Searle is concerned by the dominance of functionalism in cognitive science and with the mutual relation between this functionalism and AI. His 'Minds, brains and programs,' (1980) raises significant issues that we will set out later. Second, and mirroring this, another subcategory of AI^{S} includes the work of speculative science, futurists and of science fiction writers that project or imagine the potentials of a human (animal/alien) equivalent AI and then of super-AI. Though this latter subcategory is mainly concerned with what AI is not yet

(and may never be), discussion in terms of it already has material consequence and some focus on imminent prospects. This is because AI^S encompasses legal and regulatory discussion and development.

One of the more prominent current examples of the latter subcategory of AI^S is provided by the work of the European Parliament Committee on Legal Affairs Commission on Civil Rules on Robotics. Since AI^S focuses on intelligence in terms of what it is and what it may be, it raises issues in terms of 'autonomy'. Clearly, matters of autonomy are immediately significant for the legal status of AI. Once an AI becomes a seat of *decision-making* it becomes a source of concern regarding its material consequences. Here, there is some ambiguity that glosses over the difference between a *locus* (site) of decision-making and a *source* of decision-making, with the former tending to inform how 'autonomy' is conceived. In any case, since decision-making can be programmed, an AI^S set of legal concerns need not wait on any demonstrated extensive list of all imaginable AI^S characteristics. 'Autonomy' creates a legal issue regarding liability for actions, harms etc. since it introduces a potential break in chains of causation with reference to owners, designers and builders, and makes ambiguous the concept of 'operator' (think of a driverless vehicle, a warehouse mobile delivery unit, an adaptive-targeting drone weapon).

Moreover, since AI are also manifestly developing or changing, a legal perspective immediately invites forward thinking regarding what an AI is and may become. These are *already* germane in a way that they are not necessarily for the programmer, the data analyst, the systems theorist, the economist etc. Note, the Stanford project, like others of its kind, does not simply ignore a legal perspective but this is not its core concern, and so matters of what constitutes intelligence are marginalised. The typical default to AI^W is a matter of dominance rather than a simple denial of the concerns of AI^S. In contrast, a legal perspective on the present expressed into the future cannot marginalize matters of AI^S precisely because of the nature of law. In law, function begs questions regarding consequences and thus responsibility, which cannot evade issues of the constitution and status of entities. The current draft report prepared for the European Union Civil Law on Robotics conveniently illustrates this:

> [T]hanks to the impressive technological advances of the last decade, not only are today's robots able to perform activities which used to be typically and exclusively human, but the development of autonomous and cognitive features -- e.g. the ability to learn from experience and take independent decisions -- has made them more and more similar to agents that interact with their environment and are able to alter it significantly; whereas in such a context, the legal responsibility arising from a robot's harmful action becomes a crucial legal issue […] the more autonomous robots are, the less they can be considered simple tools in the hands of other actors (such as the manufacturer, the owner, the user etc.) whereas this in turn makes the ordinary rules on liability insufficient and calls for new rules which focus on how a machine can be held -- partly or entirely -- responsible for its acts or omissions; whereas as a consequence, it becomes more and more urgent to address the fundamental question of whether robots should possess a legal status […] ultimately robots' autonomy raises the question of their nature in the light of existing legal

> categories -- or whether a new category should be created, with its own specific features and implications as regards the attribution of rights and duties, including liability for damage […The report recommends] creating a specific legal status for robots, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons with specific rights and obligations, including that of making good any damage they may cause, and applying electronic personality to cases where robots make smart autonomous decisions or otherwise interact with third parties. (EP, 2016: pp. 5 & 12)

The EU draft report demonstrates that an AI^{S} frame of reference is already within the purview of organizations. However, it is probably more reasonable to suggest that matters of AI^{W} have provoked a partial AI^{S} response.[8] The driving force even here remains for the moment function and consequences. The draft report defers and delegates fuller development of matters relating to the status of an electronic person. For example, the profound question of whether an AI can/could flourish or suffer and whether this is *the* basis for the ascription of rights. So, whilst AI^{W} decentres what it means to be human, current tentative steps towards greater focus on AI^{S}, in terms of the AI, for and from the position of the AI, are also limited. This is partly because it is difficult to consider derivative issues of electronic persons for those persons until one has an adequate concept of the person and then some knowledge, based on realisation, of an actual electronic person -- so an *AI is AI* problem returns in a different guise (however, see Calverley, 2007). In any case, exploration of AI^{S} is a small but important aspect of the whole EU report. This pattern is repeated in terms of current legal and regulatory committee investigations in many countries.[9] AI^{S} remains mainly a matter for philosophers, futurists, and science fiction.

Shifting, adequacy and the interstitial problem

Ultimately, when AI is considered as a discourse, there is a shifting back and forth between AI^{W} as a focus on function and AI^{S} as a focus on entities (which may explicitly be functionalist). A good example of this is provided by a recent collection edited by Brockman (2015). The collection brings together contributions by leading scientists, social theorists and philosophers first published as responses to the 'question of the year 2015' on the Edge online science salon: *what do you think about machines that think?*[10] Some contributions are overwhelmingly AI^{W} and some AI^{S}. However, given the

---

[8] Gary Lea, visiting researcher in AI regulation, makes this clear in his blog on the well-known site, The Conversation. As already noted AI researchers are aware that the concept of intelligence remains ultimately ambiguous and that it might be preferable to have a 'human independent' measure; various attempts have been made to address this in terms of function that increasingly recognize aspects of general intelligence along AI^{W} lines, but do so with an AI^{S} set of concerns related to foreseeable regulatory issues: https://theconversation.com/why-we-need-a-legal-definition-of-artificial-intelligence-46796 Many individual governments are now starting to take an interest along these lines.

[9] For example, the UK House of Commons Science and Technology Committee has produced a series of reports on Robots and Artificial Intelligence. These invite expert evidence, and this includes from prominent figures concerned with AI^{S}. For example, Alan Winfield, Professor of Robot Ethics at University of the West of England.

[10] https://www.edge.org/about-edgeorg

nature of the question, which invites deliberation regarding the scope for 'thought', some contributions shift between the two, taking an AI^{W} approach to contemporary technology and its immediate prospects and an AI^{S} approach to more speculative possibilities. For example, the Harvard psychologist and public intellectual, Steven Pinker praises the focus on function as a means to overcome 'spiritualism' along AI^{W} lines, but adopts a speculative AI^{S} position to make the point that future AI need not be innately aggressive, since this is a masculine trait and not necessary to a non-masculine AI; a thought likely provoked by his own work on gradual progress in human civilization (Pinker in Brockman, 2015, pp. 5-8). By contrast, but still shifting, the MIT Nobel prize winning physicist, Frank Wilczek tends towards AI^{W}, but does so in terms of an AI^{S} first subcategory framing: 'What distinguishes natural from artificial intelligence is not what it is but only how it's made' (Wilczek in Brockman, 2015: p. 121).

Following this example, one might conclude that since both AI^{W} and AI^{S} concepts occur and are subject to development, and, moreover, it is possible to shift back and forth between them, that there is no problem regarding how AI is conceived. However, two problems arise. First, the existence of concepts does not entail concepts are adequate. For example, varieties of entity related functionalism may be critiqued in different ways. Second, legitimating shifting does more than legitimate both concepts (AI^{W} and AI^{S}), it tends to put aside how focus affects the way concepts operate. The existence of two basic focuses for AI, one where function dominates, and one where the nature of an entity dominates, creates scope for one conceptual focus to be more influential than another. Currently, and in many ways understandably, that is AI^{W}. So, the existence of two focuses creates the potential for dominant effects from the dominant concerns of a concept, since others are marginalized. Moreover, issues may be marginalized in ways that do not appear in the concerns of the subordinated conception, becoming rather interstitial. There may then be a link between the adequacy of *each* concept and the problem of *both* concepts. There is thus a prior problem of ontology that may be used to appropriately explore these matters. I will return to this later in terms of relational goods. At this stage I simply suggest that the existence of a juxtaposition of concepts and foci creates grounds for perpetuation of problems of many kinds. Consider this in terms of the difference between a spectrum view of intelligence and the possibility of breakpoints and emergence, and how this may render what it is to be human decentred.[11] One way in which this is important starts to become clear when one begins to think about Transhumanism (TH).

Transhumanism lower and upper case *(th↔TH)*

In an ordinary language sense, lower case transhumanism (*th*) is a portmanteau blending of 'transitional' and 'human', though one that also invokes transforming and transcending some prior limit on the capacities, abilities or typical observable life outcomes of the human. As such, the term is extremely broad, if not amorphous, but

[11] Note one might categorise some complexity theory approaches to AI as AI^{S} and complexity theory is typically defined in terms of emergent properties (though what this means is highly variable), so it is important not to give the impression that the purpose of differentiating AI^{W} and AI^{S} is to create a simple dichotomy. Rather it is to establish that dichotomisation is a tendency that affects nuance and subtlety and focus.

it is also intimately bound up with function. As transition etc. it involves change, and since changes to capacity, ability and observable life outcomes are not new, there is nothing intrinsically new to the notion of *th.* Throughout history humans have been changing, augmenting and enhancing their bodies, and also the context in which bodies are capable of achieving (and suppressing) things -- in relation to persons, roles, agency etc. Consider some of the range of means by which this has been achieved over time: artefacts (tools, machines, prosthetics, exo-tech etc.), pharmaceuticals, surgical-intervention (transplants, implants, amputation), technologically based services that create grounds for or affect human activity and so forth. Consider, in addition to 'augment' and 'enhance', some of the further language we apply to changes in terms of these means: facilitate, stimulate, extend, (re)generate, suppress (negate), mutilate, delegate, perfect… *And,* consider some of the historical consequences of related change for the human. Health immediately springs to mind: life expectancy, vitality, heights, weights, shapes, mental states etc. However, many other aspects of living can also be thought through in similar ways. For example, our relation to time in terms of how long given activities take, what activities are possible, our sense of distance as a time measured relation, our life ordering through clock time etc.

If translated into matters of social ontology, the above seems no more than a specific way of making the general point that humans live within open systems in process. Human history, the history of the human, has always been entangled with invention and innovation. Tool use is as old and older than *Homo sapiens*. However, changes, recognized potentials and speculations have made possible a particular discursive response regarding transition and transformation. Modern surgery, the prospects for genetic manipulation, and continual development of information technologies as hardware and software have provoked issues regarding interfaces, melding, mutation, and perhaps even re-embodiment and disembodiment. Here *th* becomes entangled with AI in various ways and this has been recognized as uppercase Transhumanism. Proponents of Transhumanism tend to use H+ or h+ to refer to it. However, I will continue to use TH. Within TH there are transitional humans and a potential for a new kind of entity, the 'posthuman', who will live in new kinds of societies that welcome and celebrate future AI as equivalent (and different) entities (see Regis, 1991; O'Connell, 2017). Though it has antecedents, the term Transhumanism began to be used in the 1980s. A World Transhumanist Association (WTA) was founded in 1998, and there are several different versions of a Transhumanist declaration or manifesto. Key aspects of the WTA declaration are:

> Transhumanists advocate the moral right for those who so wish to use technology to extend their mental and physical (including reproductive) capacities and to improve their control over their own lives. We seek personal growth beyond our current personal limitations […] It would be tragic if the potential benefits failed to materialize because of technophobia, and unnecessary prohibitions […] Transhumanism advocates the well-being of all sentience (whether in artificial intellects, humans, posthumans, or non-human animals) and encompasses many principles of modern humanism. (WTA, 2005: p. 1)

The core emphasis on the benefits of AI and on augmenting and ultimately transforming the human (and enabling posthumans) has outlasted the various incarnations of TH organizations (which are themselves in flux).[12] Notably, all versions incorporate a moral argument: the right of free expression (where TH and the posthuman are creative and liberating) and a duty or obligation to recognize and accommodate fully realised AI entities. Many adherents of TH also prioritise a 'proactionary principle' over a 'precautionary' one: the requirement to transcend risk through activity rather than avoiding or suppressing change because of recognized risks (such as the loss of identity, quasi-Gattaca coercive/competitive eugenic societies, worse-case scenarios of abrupt transition to fully realised AI -- a 'singularity' -- with Terminator consequences etc).[13] The emphasis on benefits and the weighting towards a proactionary principle is also associated with an 'abolitionist' thesis: biotechnology and social transformation can (and should) eradicate suffering as a human experience (by altering the capacity to experience). This thesis is most closely associated with David Pearce, one of the founders of the WTA.[14]

The exotica of TH is part of its appeal and has created public curiosity, often within Future Studies (see Gidley, 2017). Moreover, since *th* has created the grounds for TH as a discursive response it is unsurprising that the same set of possibilities has provoked an interest from other perspectives. Various media have pursued a 'the future is now' theme in relation to current forms of *th* advances in science that necessarily raise concerns regarding a TH position of advocacy. For example, general advances such as CRISPR gene editing, recent advances in synthetic blood production, experimental implants to control Parkinson's, as well as more specific pioneering work such as the successful implantation of microchip technology in a quadriplegic male enabling him to gradually recover some use of an arm (Mason, 2016). One might also note Elon Musk's recent corporate launch of Neuralink, a company dedicated to the development of a neural lace for human-computer interfacing. Clearly, aspects of TH potentials immediately invite legal and regulatory concern, and so parallel some aspects of how AI[s] has been motivated. This is a matter of (in the political sense) public interest and public concern. In the US, for example, research from the Pew Centre indicates popular misgivings about a transition from helping the unhealthy to shaping and intervening in the lives of the already well (Funk et al, 2016). In focus groups, the more participants were invited to consider the issues, the more it became evident that society was underprepared to deal with any consequences because of a lack of public deliberation and informed awareness (Rainie et al, 2016).[15]

---

[12] The WTA is now HumanityPlus: http://humanityplus.org
See also Transhumanity: http://transhumanity.net

[13] For the proactionary argument see Bostrom and Ord, 2006. At the extreme this becomes a reversal of a posited 'magic-in-the meat' position (see critique of Searle later). For an account of future risk from super intelligence see Bostrom (2016)

[14] There is a potential dangerous elision in the abolitionist position since there is a difference between removing suffering from the world -- changing the relations of the world -- and removing the capacity to suffer from the human; the latter does not mean ills are removed merely one's capacity to experience them. This inscribes an unpleasant potential: create inhumans because we act 'inhumanly', perpetuate harms because we are something other than human.

[15] The Pew focus groups were designed to cover significant subcategories of US society but each was constructed with an internal similarity of members in order to expedite free flow of conversation. In general participants expressed views sharply at odds with a TH proactionary approach and emphasised the need for caution and intervention beginning from a 'first do no harm' principle for the human

Given the growing attention above, various academic disciplines, and notably ethics, have oriented on *th* and TH. Again, there is some crossover here with AI, notably AI[S]. For example, there are now Professors of Robot Ethics, such as Alan Whitfield, as well as Bioethics, such as Robert Sparrow. Ethics' disciplinary interest and crossover also extends to critique or defence of TH (for an initial range see Sandel, 2007; Cabrera, 2015; Clarke et al, 2016). TH has also attracted critique from other academic positions. Though TH refers to posthumans, TH can also be differentiated to some degree from a set of humanities and cultural studies-based social theories collectively referred to as Posthumanism. These tend to be critical of TH, locating it as hasty valorisation of novelty and fantasy that does not pay due attention to feasibility or to the social theory basis of society.[16] Many Posthumanists are still engaged in rethinking theory as a necessary precursor to any emphasis on remaking society or understanding the human (see Badmington, 2000; Herbrechter, 2013; Wolfe, 2009; Braidotti 2013). The sources for this rethinking range across Butler, Deleuze, Derrida, Foucault, Latour and Woolgar, Haraway, Luhmann and many others.

Other essays in this collection have more to say about the inter-connections between TH and Posthumanism. For our purposes, it is sufficient to note that the main threads of Posthumanism take a different approach to decentring the human than we introduced earlier in this essay, and this needs to be distinguished here to avoid any confusion. Literary theory, cultural studies, post-structuralism and postmodernism emphasise the entanglement of knowledge and power and tend to associate the Enlightenment and humanist tradition with the uncritical projection of universals that are actually expressions of oppressive and marginalizing particularities, as well as sources of dangerously posed discourses of *scientistic* science that foster harms. As such, decentring the human is seen as an important theory move in opposing problems of gender constructs, ecological destruction and so forth. Many realists have great sympathy with the intent, but are sceptical regarding the ontological implications of subsequent theory and critiques.[17] There is some crossover here with issues already explored in the five volume Centre for Social Ontology working group Morphogenic Society project (see Morgan, 2016). Realists argue that the new materialism, vitalism, actor network theory etc. replace one set of problems with another set (flat ontologies, lack of adequately explored differentiation, conflation of particulars as universals in epistemology with essence and kind in ontology etc.).

(based on unintended consequences), and with due attention paid to preventing the exacerbation of current inequalities based on privilege enabling the few to pay for augmentations and changes that put them and their descendants apart (noting however that society is already unequal and that technology might actually allow for equalisation, depending on how it inhered in society). The responses also mirrored the idiosyncratic US suspicion of big government preferring some more cross-social means of oversight of technological change (see Rainie et al 2016).

[16] For an early feasibility critique see Nordmann, 2007. Note that realist critique seems to logically trade on the irreversibility of TH transformations (so something is harmed or lost without full consideration of what is harmed or lost). However, reversibility may change some of the force of argument -- since experimental TH to create mutable entities may carry different force of argument (and is a staple of space opera sci-fi utopia, such as Iain Bank's Culture novels.

[17] One might also note Steve Fuller's *Humanity 2.0* (2011) here. Fuller raises many important issues (especially the moral horizon of the human) but does so in terms of his usual social epistemology. He sets out how discourse disputes and makes ambiguous science in society but ultimately provides no definite position regarding what it means to be human or what the prospects for humanity are (or should be). Though thought provoking, the work explores evasion evasively (see Morgan, 2013).

The initial point we have made in this essay is that an AI^{W} functional approach decentres what it is to be human or, to be equivalent in terms of possible *essential* characteristics, and that an AI^{S} entity focus may not easily resolve this. Clearly matters of kind and essence are problematic for Posthumanism based on its theory sources, and for TH based on a combination of optimistic emphasis on the benefits of transformations and a proactionary principle. Again, others have more to say about this. My concern is with problems of social ontology that may become ingrained or interstitial based on how AI has been conceived. As such, I now move on to Turing and Searle to consider the 'sophisticated origins' and implications of AI^{W} and AI^{S}.

Turing, AI^{W} and AI^{S}: finding a question that can be answered

In his seminal 1950 paper 'Computing, machinery and intelligence,' Turing clears the ground for a dominant AI^{W} focus on function, whilst also setting the scene for AI^{S} sub-categorical concerns. Turing's point of departure is: can machines think? However, for Turing, the ordinary language sense of this question is too ambiguous and this impedes any satisfactory answer. As such, what is required is a substitute question that can in principle be answered. Specifically, could a machine provide responses indistinguishable from those a human provides, and so pass for human? The substitution takes the guise of a thought experiment in the form of a test, the 'imitation game'. Turing describes the game initially as one played by 3 people: an interrogator (C) and a man (A) and woman (B). C is in a separate room and communication is via some medium (another party or by cards, teleprinter etc). The interrogator (C) is unaware which of the two others is the man or woman. For C they are merely x and y. The task for C is to decide which of x and y is the man and which the woman. The man (A) is given the task of confounding the interrogator's (C) attempts to identify which of x and y is the man and woman, and the woman (B) is given the task of helping the interrogator. However, neither A nor B can simply state who is who.

Turing then proposes that a machine take the part of A, and by machine he means a digital computer of some possible future variety. This digital computer is a machine designed to carry out operations that could be done by a 'human computer'. A human computer in these terms is one that would be 'following fixed rules' and without 'authority to deviate from them' (Turing, 1950: p. 436). The digital computer (which stands in the place of AI) is then described as an extrapolation of contemporary technology: a technology that follows fixed rules where the technology is a combination of a store of information, an executive unit that carries out individual operations in a calculation, and a control table of instructions (a program code). Turing describes this digital computer as for-all-intents-and-purposes a 'discrete-state machine'. That is, one that follows rules and shifts from one definite state to another, which unless error occurs is ultimately predictable (in a basic 'Laplacian' sense). It thus has a clear set of input-output pathways. However, given that one can program a digital computer to fulfil any function that can follow this procedure one can describe digital computers as '*universal* machines' (Turing, 1950: p. 441). Turing then suggests that a digital computer with a sufficiently large storage capacity and processing speed could in principle play the imitation game.

Read carefully the imitation game is clearly a test of functional equivalence. The interrogator C is simply being asked can you distinguish between two entities in relation to tasks and indicative characteristics. Though the range of tasks can be wide the underlying structure of them for the original argument is highly circumscribed: a set of rule following actions. However, there is more involved since Turing's underlying argument can be differentiated from the initial underlying structure of the game he sets out. This becomes clear based on the examples used and the way some of the 9 counter arguments or objections to the test are addressed.

The underlying structure of the imitation game requires a machine to follow a program that enables it to simulate the responses of a human. The human point of reference discussed as equivalent in the argument in the paper is a human computer who does not deviate from fixed rule following behaviour. However, the human responses actually illustrated for the game are wide-ranging and include more naturalistic responses. They do not focus only on a human engaged in unequivocally *fixed* response answers to given questions. For example, they are not closed-end yes/no issues or simply matters of a human calculating. Clearly this would be too narrow to satisfy any reasonable form of imitation test. Significantly, then, the statement of a range requires an intuitive leap or inference that a *discrete-state* yet 'universal' machine can at some point in the future play the imitation game in a way that can answer the possible range of questions. At a minimum, it must be able to address Turing's illustrations in the paper. Most notably in setting the scene for the game: 'C: Will X please tell me the length of his or her hair?' If A is X: 'My hair is shingled and the longest strands are about 9 inches long' (Turing, 1950: pp. 433-434). And then in the following section when identifying different 'specimen' question forms: 'Q: Please write me a sonnet on the subject of the Forth Bridge. A: Count me out on this one, I never could write poetry' (Turing, 1950: p. 434). For Turing, passing such a test is sufficient to answer the question: can a machine imitate a human? However, there is more to it than this, since Turing is reasonably confident that a machine *will* eventually pass the test and in introducing possible objections he states: 'I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted' (Turing, 1950: p. 442).[18] He situates this claim as 'conjecture', but the impression conveyed is important. The claim is made based on the digital computer, the human computer and then the intuitive leap. For Turing, the current impediment to playing the game is processing capacity and speed.[19] The inference is that it is based on future technology following similar lines and with reference to the game that it will be reasonable to claim that a machine thinks (*despite* that the game is about imitation).

---

[18] An alternative argument to the one that follows is that the 'use of words' changes, and so thinking is defined differently, such that a digital computer and human think without addressing the points I make. This, however, does not help Turing's position, since it relies on semantic incoherence as a solution to substantive incoherence of argument, and so replaces one problem with another, where the new problem fails to respond to the original issue: it merely repositions and evades the issue.

[19] 'I believe that in about fifty years' time, it will be possible to programme computers with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning' (Turing, 1950: p. 442).

Now, consider what is involved in Turing's position. The point of departure is *equivalence* between a human computer and a digital computer, but the language use of the former is ambiguous once the argument starts to be extended. There is semantic slippage from the claim that a human can compute and this follows fixed rules or procedures (whose archetypal form is calculation according to formulae) to the implication that a human computing is equivalent to a human thinking. Turing's argument claims a digital computer following fixed rules can achieve imitation and also (as confident conjecture) *implies* that the technology (presumably with reference to the game) will settle the issue of whether digital machines think. For this line of reasoning to be plausible it must rely on the assumption that a human thinking and a digital computer computing are equivalent. There is thus a logical substructure that encourages both an AI^{w} focus for future AI researchers (a concentration on function where the problem of the entity will take care of itself based on technological development) and (beneath the conjecture caveat) an AI^{s} first subcategory claim that can easily become a conflation: the human mind is equivalent to a computer, so the human mind is a computer. This latter form also trades on the inference that a human mind can be reduced to input-output procedures (and so is also overwhelmingly about function).

At this stage it is important not to traduce Turing, but rather to highlight the problems of his lines of reasoning. Turing's claim implies equivalence but he introduces further caveats and considerations in addressing 9 objections. Three are immediately relevant here. Turing acknowledges that a human nervous system is not a discrete-state machine but argues that if a machine can play the game and thus be indistinguishable from a human then this is irrelevant. The implication is thus that equivalence is not identity of constitution but similarity of outcome. In the context of the claim about machines thinking, the implication is thus that behavour is the significant locus that allows the inference that a machine can think. If one reverses the line of reasoning then the implication is also that the internal operation of the human in the act of thinking is conducive to the equivalence of outcomes and equivalent status (both the human and machine 'think'). The 'difference' thus seems to make no 'difference'. However, Turing also notes the objection that the game is not a test of consciousness, since the digital machine is not required to (or demonstrate that it actually can) know what it is doing, or feel emotion. This seems to indicate that Turing is simply claiming that the test is purely a matter of simulation based on the game.

However, the very point of the game is to replace the ambiguous ordinary language variety of question (can a machine think?) with an operationally answerable question (can a machine play the imitation game?). It is this that underpins Turing's claim that in the future it will be permissible to state a digital computer can think. For this to be so, one must, therefore, invoke the underlying assumption that it is meaningful to substitute the latter question for the former, which in itself can only be meaningful as an act if one assumes the substituted question *bears on* the original one. The context and purpose of the game thus create implications that shape the subsequent caveat: 'I do not wish to give the impression that I think there is no mystery about consciousness' (Turing, 1950: p. 447) in terms of the subsequent statement: 'But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper' (Ibid). One might infer then that the 'mystery' can be answered -- Turing does not say. However,

he immediately moves onto the possible objection that there is no evidence that humans follow laws of behaviour and actual activity is variable in terms of specific conduct. Here, he notes that general laws of behaviour may exist that condition the scope of variability in specific conduct. Turing seems to be implying here that general laws inhere in the human thinking and so equivalence is an interior matter of rule following behaviour rather than merely exterior circumscribed equivalent outcomes. There is, therefore, grounds to infer based on Turing's argument that a human thinking and a digital computer thinking are (or could in the future be) doing the same thing.

It remains the case that Turing's position is sufficiently underdeveloped to allow different interpretations. As a consequence, there is a line of AI^{S} subcategory work that explores Turing's imitation game in terms of what he actually intended, including whether he intended the game to be behavioural in its implications because of its functional focus (for early examples that assess the debate see, Millar, 1973; Lassegue, 1988). This notwithstanding, the point I want to emphasise is that the direction of argument Turing pursues in terms of key objections follows a pattern. He affirms the relevance of the imitation game as a valid test and does so by orienting on the significance of behaviour, function and equivalence. This enables a slide in the argument such that equivalence is a matter of function, which is suggestive of more than mere function: function becomes the significant indicator of 'thinking'. So, whilst there are manifest tensions in the way Turing reasons, his argument conveys an impression regarding 'thinking' relevant to (and encouraging) both AI^{W} and AI^{S} despite that the game is about imitation.[20] Since it has given encouragement to both, the original problematic created by Turing is thus multiply suggestive, and so ambiguous in its particular implication, despite that it is constructed to enable definitive consideration of what Turing considers would otherwise be too amorphous a problem. This returns us to Turing's intent. Turing intends to find a question that can be answered that can stand in for an ordinary language approach to: can machines think? However, one can reasonably ask: 1) does the form of the imitation game argument *as constructed by Turing* actually provide grounds for concluding that the game can be played effectively by a digital computer? 2) is the new question actually an appropriate substitute for: can machines think? Exploring the former highlights the tensions in Turing's position, which have contemporary significance, whilst identifying the latter provides an entry point to Searle's approach.

Playing the imitation game, substitution as seduction

The question: does the form of the imitation game argument *as constructed by Turing* actually provide grounds for concluding that the game can be played effectively by a digital computer? creates multiple grounds for dispute. The original imitation game

[20] This impression is also created by Turing's reply to the 'solipsist' implications of the argument for consciousness where he states that a sonnet writing machine capable of demonstrating opinion would be unlikely to be described as merely 'signalling' (see Searle on the fallacy here). Note: it has also been pointed out that the 9 objections mainly rely on emphasising the lack of evidence to refute Turing's position i.e. absence of proof is not proof of absence. However, such arguments for the negative have the general property as argumentation structures that they contain no evidence for the argument, and as *reductio ad absurdum* this allows any non-evidential claim to hold (particularly those for which no evidence seems likely to be ever forthcoming). It is arguable how far this actually applies to AI.

requires an A and B as x and y to either seek to fool or aid the interrogator. This is an open-ended strategic problem of context, even though it is stated as two roles with a master directive for each role (in addition to the role of C). There is a significant difference between a codified response to a specific question within this remit and the strategic narrative that emerges to express that remit. A fixed rule approach must, therefore, solve the problem of *strategic conversation* rather than merely syntactic and semantic consistency of an individual response. The terms of the game thus conceal the problems of complex improvisation and the naturalistic feel that such a conversation must convey. Turing does not resolve these problems, since his actual examples (quoted previously) focus on individual or single responses rather than strings or pathways of interactive *dialogue*.[21] His first example (hair) seems to reduce questions and answers to a simple problem of logic where a game is deductive elimination (e.g. x is not B because of answer z). His second example (the sonnet) is simply a form of evasion. However, the former example as human conversation could quickly become confounded by *non sequitur*, confusion and ambiguity as characteristics of the conversation, and so a deductive approach as coding would find this difficult if not impossible to cope with. Coding responses that sought to rectify the problem by putting the conversation back on track would immediately strike an interrogator as non-naturalistic, creating suspicion likely leading to failure in the seamless substitution (machine for man) aspect of the imitation game. The very name of the game trades on a conflation of two different purposes: imitation qua man/woman for any player (implicit in the different role remits of A and B) and substitution of machine in one role. In the case of the second example (the sonnet), if codified responses to such questions are all in the form of evasions, rather than either demonstrated ability or qualitative opinion, then there is another cumulative effect of suspicion. The interrogator will become suspicious as to what kind of entity they are dealing with. This is a double hermeneutic problem once the Turing test becomes a matter of public knowledge (as it would over the rest of the century).

Both examples highlight that an interrogator is an *interlocutor* within a dialogical open process. As such, it is not clearly established that *the* Turing test can be passed based on the basic foundations of technology as stated, and extrapolated from, *by* Turing: a discrete-if-universal machine. Multiplying discrete functions is a confusion of what universal implies, since it indicates universality is merely additive. This is a problem that continues to dog contemporary AI technology. There is a difference between operative efficacy in a task and navigating seamlessly between tasks (AI researchers typically refer to this in terms of the specific and general intelligence problematics). For our purposes, there is something analogous in conversation since it is the way one adapts *and* contributes that signals constructive appropriateness. At root there is another problem of emergence here (to add to the issue of spectrum intelligence).

Moreover, careful consideration of the examples and the problem of dialogue indicate that what it means *to pass* the test is also disputable. The purpose of the test is clearly stated, but the specific design of the test and the context in which it is applied are not fully developed by Turing. As such, he does not address the issue that coding could be developed in accordance with how the test is specifically operationalised. So,

[21] Though one might refer to his objection based on consciousness point here.

'passing' the test can become a matter of passing *a* test in the context of gaming the game or substituting some other version of the test. This became a matter of some controversy in June 2014 when the 'chatbot' program Eugene Goostman persuaded 10 of 30 judges (33%) at a Royal Society organized AI event that it could pass for human.[22] The 2014 Royal Society event was not recognizably Turing's imitation game. It was a 5 minute keyboard based interaction between judges and the program, where the judges were asked whether they could distinguish a program from a human. The event set a threshold of success at persuasion of 30% of judges. The Eugene program simulates a 13-year-old Ukrainian boy, and so the form of the program creates limited expectations for the range of interactions and builds in anticipations of errors, evasions and inconsistencies that then become 'idiosyncrasies'. This is quite different than simulating a fully operational adult meeting core norms within any given yet open-ended socio-cultural milieu over an extended duration. Passing the Royal Society event test thus quickly became a matter of what kind of test was passed based on what kind of coding. The best-known AI event is the annual Loebner competition, which has offered a cash prize of $100,000 since 1991 for a program *fully* indistinguishable from a human, and a smaller prize for the best entry of the year. As of 2016, no program had won the $100,000.

However, dispute regarding test design is not itself a decisive refutation that a digital computer can be programmed to successfully play the imitation game. What one can state is that Turing does not establish that it is possible. However, again it is important not to traduce Turing. His specific development of the form of a digital computer is as a discrete-state machine with definite input-output relations. However, in discussing objections he considers the possibility of a 'learning' machine capable of 'induction' (not abduction/retroduction). He makes no attempt to articulate how this might be constructed, nor is it central to his argument, but he does at least introduce the possibility. Chatbots and related 'AI' technologies are constantly developing and the use of big data analytics drawing on a huge pool of conversation and communication creates the possibility that an effective digital computer could draw on blocks of similar responses from similar situations, and so simulate naturalistic language with a level of apparent sophistication that an interlocutor would deem appropriate. This possibility falls under the remit of 'learning' programs. The implication is that it is, as Turing claimed, only (though perhaps mainly) processing capacity and speed, and time (time for AI to 'learn', time for coding to develop as problems are identified and solved) that stand between the digital computer and successful playing of the game. There is also a convergent technology argument here, since imminent developments such as quantum computing offer the possibility of significant leaps in processing capacity and speed (if so then Moore's law does not confront the impending limit entailed by non-quantum processing). The inference drawn would then be that chatbots such as Apple's Siri and Amazon's Alexa will *become* or have descendants that are increasingly naturalistic in their interactions (pushing past momentary embarrassments, such as those created by Microsoft's Tay).[23]

---

[22] See BBC coverage: http://www.bbc.co.uk/news/technology-27762088

[23] Recent research published in *Science* indicates that though it may be possible to make chatbots more polite, since learning is based on big datasets in which meaning is embedded, AI faces a deeper problem of absorbing pre-existing human socio-cultural bias based on the way language is associated and used:

Clearly, technological change is altering what it means to talk about 'AI' and the terminology across the field is likewise changing as new fields propagate. AI research has moved on from simple discrete-state input-output concepts and approaches, and Bayesian or Boolean solutions. There is a heavy emphasis on quantifying 'uncertainty'. There is also increasing use of the language of 'complexity' to describe AI.[24] The Stanford report, for example, makes much of these as the cutting edge of the field.[25] At first sight, there would thus seem to still be some credence in the way Turing's approach seems to have cleared the ground for a dominant AI^{W} focus on function. AI will be what AI researchers do and in so far as AI become capable of passing the Turing test, what AI researchers do will ultimately (if later than Turing anticipated) fulfil the expectation that 'one will be able to speak of machines thinking without expecting to be contradicted' (Turing, 1950: p. 442). However, consider how this claim is positioned. Both 'thinking' and 'learning' are deeply ambiguous and contestable terms when applied to an AI, and arguably the shift from discrete-state as a *definite* input-output relation to more contingent approaches is a change of scope not of form.

One might argue that it is the extension of terms under ambiguity that underpins some of the difference in how argument is positioned and claims are now made. For example, one of the major innovations in current AI is 'deep learning' using artificial neural networks (ANN). ANN are described as software simulations of neuron connectivity (Economist, 2016).[26] That is, they are multiply layered sets of 'neural units' creating multiple dividing points for direction, as processing, from some given input to some output. The sophistication of the system or its capacity for difference and range is based on the number of layers, the 'depth', in the structure. What the system is directed to can then (currently) be set up in three ways expressed as learning modes: 1) supervised learning (a network system is fed an example dataset that exemplifies what it is intended to achieve, such as spam identification) 2) unsupervised learning (a network system is fed an example dataset and is set up to look for patterns, clusters anomalies in the data, which then become the specific

'We show that standard machine learning can acquire stereotyped biases from textual data that reflect everyday human culture... stereotypes and empirical associations, has long been known in corpus linguistics… since we performed our experiments on off-the-shelf machine learning components [primarily the Global Vectors for Word Representation (GloVe) word embedding], we show that cultural stereotypes propagate to artificial intelligence (AI) technologies in widespread use.' Caliskan et al (2017: p. 183). This is in addition to related problems that commercial chatbots are often designed with female voices, which are considered, non-threatening and submissive (drawing on and creating gendered effects).

[24] In realist critique there are clear ontological problems involved: the shift from simple deterministic input-output to modelled/programmed defined ranges of reaction and response produces a problem of probabilistic framing that does not transcend determinism but merely resituates it (and a language of quantified uncertainty cannot disguise this). Note also that complexity theory claims as a key component 'emergence'. One might categorise some complexity theory as AI^{S} work, but there is also a tendency to use emergence loosely to focus on functional efficacy as repeated outcome achievement (rather than differentiate events and consider emergence as a property of an entity as a source of causal power distinct from outcome). *Inter alia*, complexity sits awkwardly with spectrum claims.

[25] Revolutions have been heralded before, see Churchland and Churchland (1990)

[26] So, there is an immediate issue here since simulation of neurons is a claim that the operation is neuron-like, though once one starts to consider the actual structure of the technology rather than the claim made then it becomes clear the statement of neuron-like owes more to metaphor than substantive evidence.

output within a broader data-defined remit, such as fraud patterns in insurance claims) 3) reinforcement learning (a network system is fed an example dataset and refines its behaviour based on rewards as feedback to achieve goals, creating a simulation of 'do what works best in situation x', such as playing and winning a video game).[27] In all three cases the key innovation is that the network progressively refines the weighting between connections, and it thus fine-tunes the network system. The more data the system has to work with, the more layers to the neural network and the more simulations run, then the more effective the system becomes, over time and in real time, subject to processing capacity and speed.

Since 2012 there have been significant advances in ANN AI.[28] However, if one decodes the language of learning being used then it is about refining a system. This is termed training and/or learning because ANN does not depend on precise coding of every possible situation as an 'if y then x'. Rather than definite input-output relations one now has defined relations of inputs and outputs, but still a focused system that is all about achievement of some goal. Clearly ANN has scope to be more flexible than the coding that Turing was working with, but as yet a barrier still exists based on transferring between different specific functions (the problem of 'general intelligence'), since this still requires reconstruction of the system. Moreover, systems still require specification through function in order to exist at all. A human, arguably, is not reducible to specification qua function (and so as a being is unspecified in this sense). There is no 'I am function' for the human, as a restriction on construction and existence. Moreover, if one places the potential of ANN in the context of the Turing test, and considers the communication milieu then arguably any 'learning' AI is using language, it is not in the ordinary language sense, a language user.

The fundamental question is can one be intelligent or learning or a language user if one lacks consciousness, self-consciousness or awareness? One can simply loosen the use of the terms by extension, and trade on ambiguity. However, if one lacks consciousness etc then in what sense is it semantically appropriate to use terms that attribute understanding to an entity? At root, intelligence requires one to make intelligible, and so forth. In the case of chatbots, they can be more or less naturalistic and so more or less effective in simulating authenticity, but this does not in and of itself change the status of the chatbot in terms of the imitation game, *unless functional efficacy confers the status of thinking or unless something additional has occurred that is not yet demonstrated about the entity in question*. This returns us to the central problem created by the very existence of the imitation game. Turing replaces the question, can a machine think, with the question: can a machine play the imitation game? As we have noted, this only makes sense if we assume the latter bears on the

[27] The Economist article notes that DeepMind's AlphaGo system uses two deep neural networks, a reinforcement learning network and a random sampling network. One throws up possible moves that the other then play tests. It was this that enabled the system's much publicised achievements in the game Go.

[28] In general, 'deep learning' programs are capable of 'recognition' (objects, audio, speech etc) and have multiple applications; *inter alia* one might also note algorithmic game theory creates decision making matrices where rules can be adjusted.

former. However, the replacement invokes the second question we identified in the lead in to this section: is the new question actually an appropriate substitute?[29]

Based on the argument *so far* Turing's replacement seems like a shift that encourages problematic foci, rather than a warrantable substitution. An 'answerable' question can be inappropriate or at least insufficient, and this can become increasingly evident as time passes and subsequent work is undertaken. So, one might argue Turing creates a point of departure that ingrains a bifurcation between focus on function (with at least implicit problematic consequences for entity characteristics - intelligence, thinking, learning) and reactions that draw attention back to entities that reconsider the nature of intelligence, thinking and learning as well as *further* characteristics.

At this point one might be tempted to say: so what? If AI^{w} is a focus on function and AI research cumulatively develops to achieve specific functions then does it matter whether an AI *really* thinks, has intelligence and learns? In a trivial sense the answer may well be no. However, in a more basic sense whether an AI *really* thinks etc *really* matters because of the many social consequences of AI. Failure to appropriately conceive of the nature of entities is to invite obfuscation and this in turn is indicative of a basic ontological omission. How we refer to AI acts back on how we conceive of the human and so has possible consequences for how we value, preserve, develop and nurture the human. This relation is not a matter of effective cause, but of discursive context, raising issues regarding causation based on distinctions between exercising a power and being an operative source of influence. Moreover, the focus on function without due consideration to entities creates a kind of residual behaviouristic presumption. Thereafter, an AI functionalism can serve to legitimate TH, in so far as functionalism lends itself to the inference that an AI is the measure of the human (allowing a move where nothing significant about the human is lost in a TH or posthuman future because of tacit equivalence assumed now, which in turn leads to complacency regarding what happens now as it affects any possible future). AI^{w} thus feeds TH. Concomitantly, the focus on function can marginalise proper consideration of what is also lost for humans through what is done on behalf of humans. I will say more about this in terms of relational goods.

As a last point here consider the *inter alia* effect of the Turing imitation game. Function is highly seductive. It can become its own self-confirming technocratic discourse following AI^{w} rationales, and, as already noted, it can inspire disciplinary responses along AI^{s} lines. It is, for example, easy to become seduced by the minutiae of the imitation game. We have provided more than 3 pages of analysis regarding how and if the game can be played. This could easily be extended (and has been: Crockett, 1994; Millican and Clarke, 1996; Saygin et al, 2000).[30] As James Moor notes in his introduction to the special issue of *Mind* celebrating 50 years since the publication of Turing's paper, 'This article is arguably the most influential and widely read article in the philosophy of artificial intelligence. Indeed, most of the debate in the philosophy

[29] Note, Turing puts aside the question what would a machine that could think *think*? This also turns out to be important once one shifts to Searle's critique (see subsequent argument and also Pinsky 1951).

[30] Note longer works regarding Turing also include analysis of Searle since the latter follows from the former.

of artificial intelligence over the last fifty years concerns issues that were raised and discussed by Turing' (Moor, 2000: p. 461).

This brings us to Searle. Searle is not the first to respond to Turing (see Pinsky, 1951; Mays, 1952), but he sets in motion many of the points I have already raised and provides the archetypal AI^S argument that contests the consequences of Turing's formulation: its seductive qualities, which invite a focus on function, a concern with minutiae and a problematic slide from equivalence to conflation, where the human mind is equivalent to a computer so the human mind *is* a computer.[31] In responding in the negative regarding this latter position, Searle introduces a particular emphasis to the framework where simulation is the basis of the substitute question: can machines think? It is based on this emphasis that interstitial problems (based on the foci of AI^W and AI^S and the emphasis of the latter) can be identified. Responding to the issue of simulation creates further grounds for argument and these have resisted agreement, despite the significant plausibility of Searle's case.

Searle: AI^S, semantic versus syntax and the failure of successful simulation

To be clear, Turing is not Searle's immediate target in his 'Minds, brains and programs' (1980). His point of departure is the mutual influence that the prominence of computerisation has had for and with cognitive science, and hence problems in the philosophy of mind. He distinguishes the use of computers to study the mind (a functional tool) from the claim that a mind and a computer are the same (they function in the same way). The former is Searle's version of AI^W and the latter AI^S. His use is thus narrower than I have previously set out for these terms. A focus on function is not really his primary concern, at least in the sense of the subsequent focus on, and consequences of, how AI develops to function in the world. He is rather concerned with the problem of functionalism, and initially with behaviourism. His aim is to demonstrate that the mind and a computer, as currently conceived, are not the same (and so one cannot claim that how a computer works explains how a mind works). In so doing he acknowledges that Turing attempts to put aside the problem of consciousness, and yet the imitation game as a simulation remains subject to critique. The point of the critique is to establish that a computer and a human mind are different, even if the superficial consequences can be the same: a successful simulation remains merely a successful simulation, unless one can demonstrate that the inner workings of both mind and computer (again standing in for AI) have similar characteristics. His focus is thus on the entity rather than merely the outcome. He clearly sets out a first subcategory AI^S position, and the argument can be located as a primary refutation of the substitute question that simulation is supposed to offer.

Searle's critique takes the form of a thought experiment. The thought experiment reverses Turing's game. Searle creates a human simulation of a computer, rather than introduces a computer as a simulation into a game to identify who is (what gender of) human. The critique is now commonly referred to as the 'Chinese room' thought experiment. A person is placed in a locked room, which contains some

[31] Pre Searle, perhaps the most notable are Block and Gunderson. Note, Block provides a prototype Chinese Room argument in 'Troubles with functionalism,' commonly referred to as the Chinese Gym/Nation argument (citizens are given instructions to phone another in a network creating a pattern of calling that replicates neuron's firing: is the collective China conscious and could it be in pain?')

material written in Chinese. The person in the room knows no Chinese and so the Chinese characters are meaningless to her. A second set of materials is transmitted to the person. This set consists of further Chinese and a set of rules in her native language (English) that enable identification of the symbols based on shape.[32] This enables the person to 'correlate' the formal symbols in the first and second sets. A third set of Chinese materials is then transmitted with further instructions in her native language. This set of instructions enables her to correlate symbols in the third batch with the first two batches. Significantly, the new instructions are rules that dictate which symbols to return to outside the room in relation to the first two sets. The person in the room is unaware that the first two sets of Chinese are designated as stories/scripts and the third, questions, and that the third set of instructions is essentially a program facilitating answers to the questions.

Searle's point is that in so far as the instructions (program) are adequately set out and followed, the person is able to transmit 'answers' that are adequate, and so indistinguishable from a native (literate) speaker of Chinese. However, the person knows no Chinese and has merely engaged in formal symbol manipulation without comprehension of meaning. The program is syntax but for the operator there is no semantic content. They have acted in accordance with a program along input-output lines, and there is no interpretation-as-translation of the symbols. This is quite different than what a human does when communicating. To emphasise this point Searle introduces an additional feature to the experiment. The person is also required to answer a parallel set of questions in English. As a native speaker, the person's answers to these questions are also adequate, and so the appropriateness of answers to *both* the English and Chinese questions are indistinguishable, despite that the former are communicative-interpretive acts with semantic significance for the person and the latter are not. Searle, therefore, concludes that successful simulation is not a test (is insufficient) to establish that a mind and a computer are the same. Successful simulation is still a failure in strong AI terms (there is a confusion of simulation and duplication -- meaning is not duplicated for the operator).

As with Turing's imitation game, Searle's Chinese room has invited many critiques and responses (e.g. Anderson, 1987; Harnad, 1987; Hauser, 1997; Preston and Bishop, 2002). This began with more than 25 brief responses and Searle's replies that appear with the original essay in the journal *Behavioural and Brain Sciences*. In tone they range from the hostile to the sympathetic. More significantly, the replies are for-all-intents-and-purposes varieties of the standard objections Searle sets out (following the format dictated by Turing) as part of the original essay: the systems, robot, brain simulator, other minds, many mansions and combination replies. Not all are relevant here. What is relevant is that this sets a pattern. Searle and others' responses are unable to decisively refute the objections to the satisfaction of interlocutors. This is despite that the more considered replies concede that there is a case to be answered. The problem is that it remains possible to place a question mark

[32] Note, the Bushou structure of Chinese symbols is conducive to shape matching for parts of Chinese characters and some English-Chinese dictionaries use this format to expedite finding the pinyin, though it is not entirely clear how this relates to the Chinese room experiment in the original argument, which is purely about formal symbol matching rather than identifying meaning based on decomposition of characters in a mainly bi-syllabic language of the type Chinese is.

against the terms of Searle's critique, and so offer alternative terms or merely consider the argument incomplete.

For example, similar to Turing's game the Chinese room provokes a 'can the game be played?' response in the form of 'can the experiment be constructed?'. However, there are limits to this line of reasoning, since the argumentation scheme status of Searle's Chinese room differs from Turing's game. Clearly, both require consistency, and both involve a claim that follows from the initial construct: in Turing's case the claim is that implications (or reasonable inferences) follow regarding thinking from function, in Searle's case that they do not. However, it is intrinsic to Turing's case that a version of the game be actually constructible, whereas it is sufficient for Searle's case that the thought experiment be conceivable. Turing's case hinges on practical (albeit future) demonstration, whereas Searle's need only demonstrate in thought that such a practical demonstration is insufficient for inferences to be made regarding thought for AI. However, it is here that dispute arises, and the terms of this dispute underpin continuation of versions of all the standard objections to Searle's argument. Specifically, what does Searle's argument assume, what does it reasonably allow one to infer, and what are the limits of any substantive case?

Dispute perpetuated through the limits of argument: designing a successful failure and the context of 'insufficient'

Many replies argue that Searle has taken the position of a part in a whole and then made inferences from the part. The Chinese room argument is in this sense 'reductive', though the problem is variously referred to as a level of analysis error, category mistake etc. Searle introduces a human operator that carries information (an operating unit or processor qua program), but this is simply one component. To take the 'point of view' of a component is to miss the possibility that it is not operatively significant in isolation, and so the characteristics denied on the basis of a part may exist based on a relational whole. From this point of view, the formal structure of the Chinese room argument described in terms of a component is designed by Searle to be a successful failure. Searle is setting up a construct that *must* fail because of the position from which inferences are made, rather than he genuinely establishes that an AI could not pass an appropriately conceived *and described* test. Thus, in terms of the systems critique, he has taken a sub-system position, which cannot actually address its target. This critique then becomes part of iterations of other objections: semantics may be a property of the system, analogous to a mind, and intelligence, awareness etc may be conceivable as potentials of complex artificial systems, where these systems may emulate neural patterns, and if embodied (a robot), and so tactile-as-experiential in the world, could have or develop to be what Searle claims the Chinese room establishes that AI cannot demonstrate (intelligence, internal semantic significance, awareness etc.).

For Searle, all these replies miss the point. The room orients on a core difference: formal symbol manipulation in contrast to comprehension of meaning. There are different terms involved and these are not synonymous (understanding, meaning, intelligence, awareness, consciousness), and so more might be said about each, but this is irrelevant or superfluous to the initial insight of the thought experiment. Expanding from a sub-system to any defined actual system, and so

altering the level of analysis, does not in-and-of-itself change the status of the claims. The entity either has or does not have the capacity to comprehend meaning (and by extension has other characteristics associated with the human). For Searle, critics do not establish that it has the capacity. So, simulation remains merely successful simulation. However, this clearly does not deter critics, since they are still able to reverse Searle's point, partly because of the limits of what can be claimed from the Chinese room argument.

The argument only establishes that successful simulation is *insufficient* to establish that a computer and a mind are the same, and that an AI can have significant mind-like characteristics. It does not establish impossibility. As such, responses continue to develop along three mutually related lines: (1) what intelligence, understanding etc are is more ambiguous, contingent, contestable, and nuanced than Searle allows; (2) Searle is missing something in terms of the actual equivalence between how a mind and computer operate (properly described the/a technology can be what Searle states it is not, and so can a mind); (3) Searle is addressing a problem that technology is progressively overcoming (and so potential is being missed by the way Searle mis-specifies the problem in parts). From this point of view, Searle's intervention has ultimately become part of the continuing discourse initially set by Turing -- not least because it reprises and so iterates the problem that dispute regarding a test design is not itself a decisive refutation that a digital computer can be programmed to successfully play an imitation game, from which inferences can then be made. Whilst one can argue Turing does not establish that it is possible to construct a digital computer for such a game, it has remained the case that perhaps one could be constructed. Searle, does more than any other to establish that simulation is not sufficient for an inference to equivalent characteristics of an entity. However, his claims have remained subject to dispute, since the primary power of the argument is based on insufficiency.

To be clear, Searle does not claim a future AI could never have characteristics we associate with the human mind (he acknowledges this to be an empirical issue). His argument is that proponents of AI have established nothing beyond formal symbol manipulation, and so claims in cognitive science that the mind is like a computer (and a computer is like a mind) should not have foundational status. If read in this narrow sense, it seems curious then that the majority of replies seek to question the basis of what is, in this context, meant only *to question* what seems an unthinking presumption about the nature of thought. The problem is that Searle does more than question in this way (his context is broader). As Searle notes, the broader issue for philosophy of mind is why would one *assert* that a mind and a computer are the same, and why would one persist and pursue lines of reasoning that first require this assertion. The context, seemingly, is a basic fallacy of reasoning that has then dogged cognitive science, which in turn affects the AI problematic: initially in the form of an overt behaviourism and eventually in a residual form (Searle, 1980, 1985, 2002, 2010). For Searle, this in turn is indicative of an odd form of dualism, where mind is separable both empirically and conceptually from the brain (a computer and a mind are the same, and so the brain is either irrelevant in itself or equivalent in its functioning to a computer). The implication is that cognitive science seems to be mis-specifying the problem of constitution and causal powers of an organic brain in terms of the problem of mind (opting for a mind is like a computer approach, which is syntactic in structure,

and fails to account for meaning as content in relation to its biochemistry), whilst AI is benefiting from the mis-specification through a unifying functionalism (what both AI and a mind are is defined by what they do).

Clearly, for this broader set of issues to make sense one needs to go beyond the Chinese room argument. One needs a philosophy of mind argument. For Searle, this has involved an explicit turn to ontology, and in terms of the broader aspects of the development of the human and of the problem of AI, a social ontology. The link here is not immediately obvious. However, consider that Searle's argument is underpinned by his general approach to intentionality. Intentionality is the capacity for 'aboutness' of the mind; its capacity to create mental states in regard of or with reference to states of affairs in the world (and this is more than just 'I intend to do', including also belief, desire etc.). For Searle, it is manifestly the case that in humans this is a consequence of the human brain as organic or biological phenomena. Awareness, consciousness, self-consciousness, understanding, meaning, intelligence and so forth are likewise biologically caused (or as Searle more often states, constituted) and involve causal powers in the world.

For Searle, a key feature of humans is their sociality. Intentionality allows for collective intentionality, defined as the capacity for intentionality to be we-directed: this 'we' remains individual, and does not necessarily require joint thinking, but always involves a referencing to what others can in combination enable or do that affects the capacity for what one is also doing, through the assumption that others are following similar mutual points of reference. This mutuality is grounded in or becomes an organizing feature of the social world built up from 'status function declarations', typically in the form of 'X counts as Y in C'. The declaration imposes a function on objects and people that are not simply performable by virtue of physical structure (they are products of the ascription or recognition of status -- a wall becomes a property boundary). From this general form, complex institutions (essentially grouped rules) develop, and this is the basis of a constructed socially reality within which one can refer to institutional facts (John is a Professor at Berkeley), where the whole is heavily dependent on linguistic representation and hence meaning. Humans are by virtue of biology capable of meaning, and it is through various related capacities that the very possibility of AI arises through technology within societies that also depends on meaning. Searle, of course, may refer to function (status function), but there is only a superficial lexical similarity between his concerns and those of the functionalism he opposes. For example, status functions are about the meaningful pursuit of living socially not the determined efficacy of completing a task (though this may be a goal of living socially). Concomitantly, to emphasise functionality rather than causal power or constitution is to mis-specify being in terms of doing, and this is a basic problem of how both the human and society are conceived.

The point to make here is that there is continuity and coherence between the different aspects of Searle's general argument, and that the combination makes sense of his opposition to functionalism. For Searle, developing his position on philosophy of mind and social ontology (beginning first with his work on speech acts) has been a life's work, and from his point of view that work augments his specific claims set out in the Chinese room argument. For critics, however, it indicates that his claims in the Chinese room argument do not stand-alone. They require commitments that are not part of the Chinese room argument. Searle essentially creates the challenge: prove

that an AI understands what it is doing. The default intrinsic to Searle's position is that *we just know* that it does not understand (based on comprehension of meaning). We know this because we designed/programmed the entity, and we know this is different than what we know about ourselves. The case of AI seems to be, therefore, an either/or issue that hinges on evidence in a way that is different than how we attribute in other cases. Those other cases are shared aspects of the human (degrees of awareness etc) one might attribute to animals, what we attribute to other humans (since it is reasonable to assume they are like us), and what we might attribute to aliens (since they may be like us). Each of these is not designed by us and so does not require some 'forgetting', which the AI case, from Searle's point of view, seems to require.

However, though hinging on evidence Searle's case is grounded on theory; that is, Searle's cumulatively developed broader position. Moreover, plausible as the focus on evidence seems, it raises the problem of what exactly would satisfy Searle in the form of proof. It cannot be mere function and so cannot be a black box behaviourist proof, but this too seems to *disallow* as much as it *disproves* a functionalist response. Since the test will be of something artificial, to critics, at the extreme, and despite caveats, Searle seems to have (at least inadvertently) disallowed any practical demonstration because he seems tacitly committed to the claim that an artificial entity is by definition synthetic and so can only demonstrate simulation. Put another way, though Searle claims that AI confuses simulation with duplication (simulating understanding is not actual understanding, so is not duplicating it) the reverse is that he resists the possibility that anything other than a brain can duplicate, and thus realise given states (and yet a synthetic heart is not a simulation-only). Searle inadvertently over-writes fallibility and future contingency via current 'forgetting'.

So, for critics, Searle has done more than he set out to do. This is a basic vulnerability that critics have developed to different degrees and with more or less sympathy for the original argument: his position takes a plausible intuition regarding difference, but requires its own assertions regarding what might be the basis of understanding; it thus involves a tacit certainty, which critics can draw further inferences from regarding prejudice or 'magic in the meat' 'chauvinism' -- the significant status accorded to the brain as the seat of given characteristics confuses significance with special or unique or spiritual or mysterious…[33] One can read Searle as hinging the difference on the nature of response: a human responds to meaning because of meaning (and so meaningfully), an AI responds to meaning because of form (and so mechanistically). The two are causally different. Meaning is not the reason-as-cause for the response of the AI (so there is no semantic just syntax). However, for critics, if there is causation then cause motivates content, so, subject to redescription, one can begin to claim that meaning is *produced* (semantics can be derived from the causal process of which syntax is a part). The question then becomes, how it is produced and what produces it, re-opening up lines of inquiry and argument that trade on complex system interconnectivity for both mind and AI, contesting the way semantic, intelligence etc. are defined and used, speculating on the basis of possible ways to duplicate (rather than merely simulate), reconstructing an argument that a

[33] Magic in the meat, derives from the science fiction writer Terry Bisson who resituates the context as a conversation between two entities who have never come across organic minds before.

robot can duplicate aboutness through sensory capacity (as experience) and so derive (a functionalist founded) meaning capacity and so forth.

Clearly, for Searle this again all misses the point. As far as we know there is no equivalent mental state in an AI *as-is*. Subtle violence seems to be perpetrated regarding our common understanding of understanding (though this too becomes a counter-argument -- since common sense plays no role in much of science and can, as convention, impede proper understanding or investigation). However, the point here is not whether one endorses or concurs with each argument, but rather that the form of Searle's case continues to provoke responses along these lines. Insufficiency, assumptions, consequences for argumentation, and then dependence on context material, provide the grounds for reasoned if not necessarily always reasonable or sympathetic responses.

And so debate continues to evolve in and around the Chinese room argument based on multiple lines of development that use it as a point of departure, trade on the limits of what it can claim, and situate this to various concerns of the critic.[34] The argument has unified critics as a point of convergence for disparate argument rather than agreement: eliminative materialists working in neuroscience and philosophy of mind (e.g. Churchland and Churchland, 1990), professional philosophers able to deconstruct the case and parse its many implications as they pertain to semantics, functionalism, alternative views of intentionality etc (e.g. Boden 1988; Chalmers, 1992; Pinker 1998; Fodor, 1992; Dennett, 2013), and futurists with agendas that entangle AI and TH (Kurtzweil, 2000).[35] As a point of convergence, then, Searle's Chinese room has not created consensus, but rather a focal point around which disagreement coalesces.

Again, this is ostensibly odd if one takes the Chinese room argument at face value: a thought experiment that places a question mark against equivalence of AI and that reminds cognitive science that the brain is significant for the

[34] It is ironic perhaps that I am pointing this out as an exercise in doing the same.

[35] Pinker, for example, stands on the opposite side to Searle as committed to a variety of functionalist information processing ('computation' and 'program') mind-AI approach: 'The mind is a system of organs of computation, designed by natural selection to solve the kinds of problems our ancestors faced in their foraging way of life, in particular, understanding and out manoeuvring objects, animals, plants, and other people. The summary can be unpacked into several claims. The mind is what the brain does; specifically, the brain processes information, and thinking is a kind of computation. The mind is organized into modules or mental organs, each with a specialized design that makes it an expert in one arena of interaction with the world… On this view, psychology is engineering in reverse. In forward-engineering, one designs a machine to do something; in reverse-engineering, one figures out what a machine was designed to do… The computational theory of mind is indispensable in addressing the questions we long to answer… the content of brain activity lies in the patterns of connections and patterns of activity among the neurons. Minute differences in the details of the connections may cause similar-looking brain patches to implement very different programs. Only when the program is run does the coherence become evident… The computational theory of mind is not the same thing as the despised 'computer metaphor.' The claim is not that the brain is like commercially available computers. Rather, the claim is that brains and computers embody intelligence for some of the same reasons.' (Pinker, 1998: pp. 21, 25-6). Pinker positions Searle's Chinese room as an argument that trades on common sense but not science, makes ambiguous the link to biochemistry of the brain for mind (in a way that information processing does not), whilst claiming that the biochemistry of the brain is core to consciousness, intentionality, awareness etc. which he never properly defines or explains (Pinker, 1997 pp: 93-96). Note, though Searle's Chinese room has become a key issue and so point of departure, it is not central to all works that explore it. Fodor (1992) is equally concerned with Churchland on meaning.

experienced/observed characteristics of the human (it has causal powers by virtue of its constitution). However, it is less odd when one considers that the strength of Searle's argument is also its weakness: it has a tight argument for a clear distinction, seems to demand an empirical response, extends to expose the assumptions that have prevented an answer he deems plausible (the mind is a computer, though this too is ambiguous if one means only computational), and is situated to an alternative theorization that grounds an answer (his ontology of human intentionality etc.). The weakness is that the combination provides for multiple lines of reasonable reply. Concomitantly, Turing's game and Searle's room are by far the most cited works on AI (partly because use exceeds a focus on AI only). In April 2017, a Google scholar search on 'Mind, Brains and Programs' returned over 59,000 results, and 'Computing, Machinery and Intelligence' more than 176,000.

Ironically, Searle's attempt to simplify and focus debate in terms of a core difference, which implicitly contests the problem set in motion by Turing of a seductive tendency to proliferate debate based on the exploration of minutiae, has produced a maelstrom of minutiae. This has had some positive consequences, since it has fostered an AI^S focus on entities, pushing functionalism to become more carefully considered regarding how and what produces function. *Inter alia*, it has contributed to formal repudiations of behaviourism (and, less constructively, for a tendency for it to become a term of abuse in discourse whose reversal is a pejorative reference to 'mentalism'). However, the positioning and influence of the argument has had further consequences.

AI^W and AI^S: resolutely unresolved lines of inquiry

We began this essay by differentiating AI^W and AI^S based on focus. AI^W involves a focus on function that tends to set aside entity status. AI^S involves a focus on entities that decomposes into two subcategories: a focus on the equivalence between human and AI, with a focus on significance for the human, primarily within philosophy of mind, and a mirroring focus on the status, characteristics and potential of AI entities. Searle's Chinese room may be about AI but it is intended to demonstrate something about the human. It is in this sense first subcategory AI^S. However, as we have set out, Searle's argument has not created agreement. It has become the point of departure for reasonable disagreement. If one works backwards through all the material we have considered so far in this paper it should be clear that disagreement is not mere formlessness. The *grounds* of disagreement are concerned with function. Moreover, in so far as responses to Turing and to Searle have contested functionality, much of the debate has concerned ways to preserve functionalist ways of thinking (no irony intended) about the problem of AI and of the human (mind). Much of the development of argument has been neo-functionalist, and that which has not been has been about the limits of critique (sufficiency of argument via games, tests, thought experiments) of what is also functionalist by focus. Despite Searle's intervention, there has been no immanent critique that has decisively shifted the terms of debate. Though it is not false to say there has been an ontological turn in philosophy of science and social science in some ways and to some degree, one cannot reasonably claim that explicit ontology or social ontology are standard points of departure for the problem

of AI. On the contrary, that Searle's position is situated to an ontology and social ontology is used against his Chinese room argument.

There is, of course, nothing intrinsically illegitimate in critique that questions the relations in a situated argument. However, there is a danger that one conflates critique of the specifics of the way an argument is situated with a basic general problem that an argument is situated at all. This is particularly important in locating Searle's work. All argument-as-claim ultimately involves an ontology, and so the alternative to Searle's argument should not be a refusal of ontology but an explicitly addressed ontological argument. Read appropriately Searle's Chinese room argument is not merely augmented by his ontology, but specifically rooted in ontological issues. It begs questions of functionalism regarding the constitution of entities and their causal powers. As such, much of the perpetuation of functionalism in spite of Searle's argument is a refusal to engage at the level of ontology, whilst pursuing tacit ontological issues. This has had observable consequences.[36]

Function has normalised as the overwhelming consideration in and for AI. Of course, noting this phrasing can seem superficially ridiculous. How could AI *not* be concerned with function? However, the nature of concern flows from what is considered and in what ways. Focus does not necessarily create clarity, but rather potentially adverse normativity. The very existence of AI^{w} is a deferment of the status of entities that presupposes the ineluctability of AI as technology. There is an 'AI is coming and we must cope' that decentres the seat of decision making, as though human choices were not dictating whether and what kinds of AI develop and are adopted. 'Cope' becomes 'let's get on with it' as though the basis of function was settled. This has a 'meanwhile' or *inter alia* context. AI^{s} may not have settled anything, but the basis of non-settlement invites a focus on function, and typically presumes a functionalist frame of reference. So, dominant aspects of the concept of AI^{s} are at least associated with the general pervasiveness of AI^{w}. As such, the bifurcation between these two that I previously referred to is not without mutuality, and this is important, since it is because of mutuality that some issues or foci or ways of conceiving are marginalised, inadequately developed or become interstitial.

As already noted, if functionalism dominates then the problem of being becomes a problematic of doing, which in turn can become a problem of efficiency. The human 'doing', as tasks, becomes a taskmaster mastering our sense of what the human is. This is sociological rather than purely philosophical (involving the positioning and relative power of ideas, rather than just the substantive content of those ideas). A focus on efficiency, for example, may absorb the social context that dominates and expresses (represses) intrinsic aspects of the human. Efficiency is a technical term but also a shaped value; it is referenced to the socio-economic

[36] Though Searle is not concerned with the problem of AI^{w}, as we have set it out, his own AI^{s} argument refuting a functionalist (behaviourist) variety of strong AI, with reference to philosophy of mind, does provide an argument for how a split in focus might arise because *both* sides of the split can share a concern with function. This then can operate ideologically and so be influential far beyond any issue of 'one thing directly causes another'. It is not difficult to understand how a program-centred view of mind in cognitive science can operate to decentre a concern with what is specifically human. It is not difficult to understand how this way of thinking about thinking (and other characteristics) can affect processes of change. Clearly, one can recognize basic entanglements here with TH, specific and general. The various exchanges between Searle and the TH futurist Kurzweil are the most obvious manifestation of this.

conditioning of the system in which values inhere (principles of capitalism that affect what humans do in the name of efficiency).

It may well be the case that ethics are a profoundly important part of AI^S, but they are not central, controlling or most important as a source of consequence in the world *because* of AI^S. Furthermore, if the dominant AI^W discourse starts from 'cope', then the power of ethical discourse to shape the social consequences of AI (and also then TH) is subverted. It is also obfuscated in so far as functionalism may inform the concepts of entities to which ethical analysis is applied. Moreover, consider the way language use has changed as and for AI. Despite critique of the extension of meaning regarding thinking, learning etc, meaning has still been extended, and so has ultimately been appropriated. Turing claims and Searle concedes that a future AI may be thinking, intelligent and so forth, *subject to how these are conceived and subject to some empirical test or demonstration.* In the meantime, the language of what has not yet been incontestably achieved has been colonised and so normalised.[37] A future that may never be is already here in terms of the language we use regarding what artificial entities *do* but which ultimately expresses an is (a be): 'intelligence' is an accepted everyday associative term qua AI, according to everyday referential communicative acts between humans AI do 'learn'... Socialisation through language use is already occurring around function. The world we live in is thus drawing us into a future we are constructing as ineluctable via language that contributes to a 'we must cope' mentality.

What follows from function:
Interstitial problems and ontology as critique

The first point to make here is that ontology itself has become interstitial. It is not the typical point of departure for AI^S and, by virtue of its focus, is not a primary concern for AI^W. Noting this is by no means to denigrate the sophistication in its own terms of work that has been done. Nor is it to invite unreasonable expectations of what ontology can achieve. Rather it is to note that all arguments-as-claims involve an ontology, and clarity here affords clarity to other and subsequent issues. This is not to suggest philosophy is unclear, or at least sets out to clarify, since analytical philosophy in particular is concerned with precision and clarity. However, such clarity is epistemic and need not be ontologically posed in general or realistically referenced in particular. Arguably, the familiar functionalism articulated via theory is imposed on reality rather than derived from it.

Moreover, clarity is situated to focus. This, essentially, is what this paper has argued by exploring the development and consequences *of focus.* Much of the seduction of Turing's imitation game puts aside ontology, inviting development of the game, whilst seeming to trade on a functionalist set of claims that then become the point of dispute in terms of what the game can demonstrate (via its substitution of questions). Ultimately, functionalism and the equivalence between AI and the human are claims about being. As such, Searle's response is an ontologically motivated reply. The Chinese room thought experiment contests the equivalence of AI and human in order to highlight that the significant characteristics of the human in relation to

[37] There is a great deal less talk of artificial stupidity.

meaning cannot be demonstrated for an AI, and yet are a consequence of the constitution of the organic brain that creates causal power in terms of mind, though Searle is wary of this language of distinction between brain and mind due to its historical legacy and connotations.

Of course, many ontologically oriented elaborations are possible, and an ontology is not assertion but reasoned argument subject to evidence, critique and subsequent critically posed development.[38] So, as already noted, Searle's argument is not immune from critique, but critique need not be the abandonment of ontology. For example, as Tony Lawson notes, Searle resists the language of emergence. According to Lawson, the subject matters of Searle's ontology and social ontology seem to require a concept of emergence to express properties for and in entities and systems (Lawson, 2016). In his reply to Lawson Searle rejects this claim, in so far as what emergence actually is, is not made clear, and what emergence explains that constitution does not is likewise unclear (Searle, 2016).[39] However, Searle's argument turns on the naming of an entity, which has an organization, and then the distinction between what follows from the organization in terms of properties, and what is known and accounted for in terms of those properties. This distinction is recognizable from early twentieth century emergentist philosophy, but creates new problems in terms of the distinctions, if the point of emergence is first-and-foremost to express the properties that would not occur without the organization, and so the properties are irreducible to the properties of the decomposed parts -– structural integrity is a characteristic of a building, consciousness of the brain, and trust of a community, in so far as appropriately constituted and active/activated. All require the organization-as-constitution to be manifest or to be possible, but not all involve an additional property that is unexplained that we deem additional *since it is so far unexplained* (as it does in the case of consciousness).[40] It is this additional unexplained that Searle

[38] So, ontology may provide clarity but it does not guarantee that matters are settled… It merely creates an additional and appropriate context of argument by recovering the traditional domain of metaphysics that much of modern philosophy eschews.

[39] For another constructive critique of Searle see Elder-Vass (2012), and for an assessment of Elder-Vass see Morgan (2014)

[40] The point at issue seems to be that additional implies cannot be reduced to, and the measure of this is cannot be explained in terms of; however, the ontic characteristic of the organization is the creation of the grounds of a property not whether in fact the property is fully comprehended regarding the organization-as-constitution. If this were so then the very moment we had a full explanation of consciousness it would be fully accounted for in relation to constitution and so would cease to be emergent by definition, and yet remain a property that would not exist without constitution. The hinge thus tacitly seems to be a difference of meaning and emphasis regarding the relation of known and additional whilst the argument itself is expressed as 'fully accounted for by x and therefore reducible to x', where *to x* becomes confused as a which *to x*? the parts or the constitution, and then the constitution as an x or a new meta-encompassed constitution of this original x – brain is not sufficient therefore brain and new x is constitution is mind… etc? In all cases constitution is the named entity that is then a source of the property, but the property is always describable as arising from the constitution. Just because one knows and designs a house to have structural integrity does not mean structural integrity does not 'emerge' from the constitution-as-combination (it is a product of…). One can of course argue about whether one wants to term this emergence or simply state that the constitution produces (where the latter still does not differentiate consciousness and structural integrity except in so far as currently known – an epistemic rather than ontic distinction), and one can argue about whether one wants to refer to consciousness and artifacts as of the same kind or category of this more general category. It is these that Searle seems to actually be contesting and Lawson does not help

associates with emergence. However, if social ontology requires a concept of emergence, it is not quite the one that Searle is rejecting. It is rather emergence because of constitution -– a reference to a range of properties *of properties* in relation to organization. What seems to be at stake initially is language use for clarity, but this ultimately creates different ways of developing similar concerns.

Lawson's social ontology overlaps with but is different from Searle's approach. It distinguishes communities (organized systems of relations expressed through rights and obligations for socially positioned humans) and artefacts (non-human though socially constituted and specified objects) as emergent social entities, and makes particular reference to language as an important emergent system. One might also note that since for Lawson artefacts are not communities, AI introduces a further potential issue for Lawson's social ontology, by virtue of what an AI may be. In Lawson's terms an AI could be an artefact that participates in language using communities, though whether one would describe AI as an artefact would then be at issue.

Lawson's is a constructive critique of Searle's social ontology.[41] But Searle's use of AI is also instructive here regarding the consequences of the sub-categorical split in AI[s]. In Chapter Six of *Making the Social World* Searle makes the case that social reality creates desire-independent reasons for acting. He does so by extending the argument set out in the Chinese room thought experiment. Specifically, he wants to contest that reasons to act only arise if one also has a desire to so act (Hume's 'reason is a slave of passion'). However, for Searle, deontic powers transcend desire or purely personal motivation. Social reality creates cross-referenced duties, obligations and requirements. In simplest form, promise keeping demonstrably creates desire-independent reasons for action. In general, we use institutions without destroying them (extending sometimes to recognizing that reproducing the institutions is necessary or important to do -- the duty to vote). We routinely suppress inclinations and modify behaviour in ways that override 'I don't feel like it right now' or 'I'd rather just do x' (we go to work, we respect property rights etc). In so far as this is so, deontic powers are a feature of institutions, and the combination affects how and about what reasoning occurs and, therefore, what we do.

---

himself here by emphasizing novelty, which seems to imply cannot be or is not known (which is different but related to cannot be predicted – the qualifiers matter as do matters *a posteriori*). Acting back upon creates a dispute in terms of organization as constitution (but even here one can argue that a house or artifact by virtue of constitution affects the durability of and environment within of the artifact – structural integrity is also a constituted causal effect for decay if redescribed, whilst material cause remains a different issue…

[41] For example, Lawson states: 'Instead of viewing individuals as materially/practically positioned as components of a totality, however, Searle, seemingly proposes a more mentalistic or representational approach. According to it individuals rather are merely 'counted' as in effect being appropriately positioned, with associated positional powers or functions. I say 'in effect' because *positions, positional powers/functions,* and *positioning* are not Searle's language.' (Lawson, 2016: p. 370). For Lawson, the development of language presupposes practices; social objects do not require contradiction in the sense an object becomes materially two things, rather that for it to be social involves a process of emergence that creates communities within which organisation, social positioning and use allows for the object to be social (one object may be many things, but it is neither contradiction nor meaningless to refer to them as *social* objects, and it is insufficient to redescribe the whole as an object with a status function - a computer may be a component in a variety of systems without being more than one computer or a computer and not a computer).

Crucially, the deontic powers of social reality only make sense if one assumes 'a gap'. That is, the possibility of *not* fulfilling an obligation or recognizing a right. This is free will, though this is not a phrasing Searle is comfortable with. Rather, he chooses to emphasise that the constitution of social reality presupposes the possibility of refusal, and so of choice. For example, an obligation is not an obligation if there is no possibility of *deciding* to fail to execute. It becomes merely stimulus-response, and this would be unrecognizable as human social reality (and the same applies to promise keeping and other variations). Searle clarifies this point by contrasting human activity with a programmed robot. The robot in Searle's construct lacks consciousness -- its response is a 'determined mechanical emission' (Searle 201: p. 136). However, deontic powers require consciousness and reflexive capacity in order for recognition of institutions to affect choice. 'I may not feel like doing x but will do it anyway…', is quite different than determination-as-compulsion in the form expressed by the behaviour of non-conscious artificial entities.

The point here is that Searle's main social ontology use of AI remains contrastive with the human, following the insight developed in the Chinese room thought experiment regarding symbol manipulation. Searle relocates his critique of functionalism, residual behaviourism and cognitive science in order to support his argument for the construction of social reality:

> The notion of a deontic power makes no sense unless you presuppose consciousness and the gap. Once you regard the creatures as like the computational models common in cognitive science, then, it seems to me, you cannot have institutional reality in our sense. You might program the machinery to resemble some of the forms of institutional reality, but the substance would be removed. (Searle, 2010: p. 137)

I by no means wish to suggest the robot argument used here is ill-founded. Rather I want to suggest that it illustrates ontology can be pursued in a variety of ways, once situated to or using the problem of AI. Other elaborations are possible. Consider, that Searle's main focus on AI is to emphasise that the brain matters for the capacities of mind we are familiar with. As such, his focus is what AI does not demonstrate regarding what mind is. Equally, however, one might ask what would be the causal capacities of a conceivable AI mind, since it surely follows they need not be the same as those of a human, if the constitution of the entity is different. This is second subcategory AI^{S}. It is contrastive in a mirroring sense. An AI entity *may* have equivalent characteristics as categorisations to the human (intelligence, understanding, consciousness, self-consciousness… intentionality), but it does not follow they will have all or similar characteristics within those categorisations of characteristics. As will become clear, this can have additional important consequences.

Constitution and second subcategory AI^{S}

One can imagine that differences create basic Nagel (1979) problems of phenomenological divides.[42] Consider that human memory is not eidetic functionality

[42] Phenomenology concerns the world as we are rather than as it is; the focus is experiential. This is a different emphasis to, but not a denial of, ontology-as-realism. Phenomenology may argue for but does

dictated by processing power and subject to cumulative 'error'. It is the drawing together of fragments according to narratives and purposes (projects). It is inherently creative and transitional; the very act of recalling repurposes and overlays in and through time. We are used to framing a problem of memory as one of unreliability of memory and witnessing, but this is a reductive sense of what memory is, referenced to a capacity the vast majority do not have (perfect recall). It tells us little about the actual significance of memory. Memory is active, and in this sense cannot be 'dispassionate' (see McGilchrist, 2009).[43] This is different than an objectivity-subjectivity dichotomy or a reality and appearance problematic (see Collier, 2003; Rescher, 2011). Active memory is a constituent in the temporality of human being; our being develops as we exist through time, and as we reconceive our past and direct it to the future.

Contrast this with AI as archetypally conceived. What kind of being is the product of eidetic functionality? What is learning and experience to an entity of error correction? Moreover, consider the difference in the grounds of being through time. Human temporality is finite; not only does it have an expectation of an end, it is experienced as an unsteady process of degeneration and bodily change in ourselves and observed in others. Leonard Cohen expresses this as:

> Everybody has experienced the defeat of their lives. Nobody has a life that worked out the way they wanted it to. We all begin as the hero of our own dramas in centre stage and inevitably life moves us out of centre stage, defeats the hero, overturns the plot and the strategy, and we're left on the sidelines wondering why we no longer have a part - or want a part - in the whole damn thing. Everybody's experienced this, and when it is presented to us sweetly,

---

not entail a radical disjuncture between reality and appearance. Nagel's famous bat argument also comments on Turing: 'Consciousness is what makes the mind body problem really intractable… [and reductionist] discussions of the problem give it little attention or get it wrong.' (1979: p. 165). He differentiates this from the Turing machine-IBM problem which are 'successful reduction' but are unlikely to shed light on the mind body problem since 'we have at present no conception of what an explanation of the physical nature of a mental phenomenon would be.' (1979: p. 166) 'The most important and characteristic feature of conscious mental phenomena is very poorly understood. Most reductionist theories do not even try to explain it… the fact that an organism has conscious experience *at all* means, basically that there is something it is like to *be* that organism… fundamentally an organism has conscious mental states if and only if there is something that it is like to *be* that organism - something it is like *for* the organism. We may call this the subjective character of experience. It is not captured by any of the familiar, recently devised reductive analyses of the mental, for all of them are logically compatible with its absence. It is not analyzable in terms of any explanatory system of functional states, or intentional states, since these could be ascribed to robots or automata that behaved like people though they experienced nothing.' (1979: p. 166) According to Nagel, if physicalism as reduction to material states is to be defended then phenomenological features must themselves be given a physical account, but this seems impossible in so far as 'every subjective phenomenon is essentially connected with a single point of view.' (1979: p. 167) Nagel's main point is that psychophysical reduction is a move towards greater objectivity by removing species-specific points of view toward the object of investigation, in terms of general effects and properties that are not simply a matter of human senses. However, experience cannot follow this pattern, since moving from appearance to reality makes no sense as a way to conceive experience (1979: p. 174).

[43] We also experience moment-to-moment incoherencies, disjunctures, fragments, flashes of images etc but we do not experience these as error we simply accept them as part-and-parcel of being and from incongruity other expressive potentials are built: humour, irony, inspiration and so forth.

> the feeling moves from heart to heart and we feel less isolated and we feel part of the great human chain, which is really involved with the recognition of defeat. (Ellen, 2016)

Brain states may be bio-chemical (involving neurons, oxytocin etc), but what would sympathy (recognition of another's - typically - adverse situation) and empathy (the experienced feeling for another's - typically- adverse situation) *be* for an entity without the biochemistry, and without first person expectation and experience of finitude, degeneration and suffering? Where would sentimentality and compassion come from and how would these be exhibited? How and in what ways would an equivalent AI socialize and be socialized vis-à-vis fellow feeling? One could, of course, seek to code or design for synthetic feeling (but again what can this *mean*?), and *inter alia* design an AI with a nervous system, and so a system for pleasure and pain. One might construct this as a component of sensuality, but what would be the general *quality* of emotion in relation to sensuality for such an AI? In a human senses, sensuality, experience and emotion are not identities, and how one becomes another is little understood (even if it can increasingly be mechanistically correlated). So, what kind of emotive states if any would an AI have or could an AI be given the capacity to have?

Moreover, consider what may be basic to difference in AI. I am aware that I have brain states, but awareness and self-consciousness have only limited access to and control over brain states (I can calm myself but cannot negate stress creation -- though some TH advocates aspire to this). But AI consciousness seems to involve code-awareness of a different order to brain state awareness. An AI may possess eidetic functionality, but it may also posses the capacity to (re)write code. As a being it would thus be potentially self-altering and the consequences of this for being in the world are difficult to conceive. Eidetic functionality and fundamental mutability of self seems a profound clash, creating a conflicted constitution of self (an inner life and introspection that confronts preservation and continuity challenges to being). Self-cultivation may share a language frame but not a real meaning sense between human and AI.[44] AI may be its own project in quite a different sense than is the case for a human (unless, of course, one aspires to TH).

Consider also what kind of experience of the world would this self-cultivating being be directed through? Self-consciousness and embodied first-person perspective are basic to a developed average adult human. An AI could have dissipated consciousness and multiple points of simultaneous perspective and experience; even if it was (partly or some of the time) embodied.[45] And in what sense could an AI *desire*? It may be the case as Searle suggests that we have desire-independent reasons for acting, but humans want things and this is subtly different than having goals. Society would be different without desire, since many aspects of how we engage with

[44] *Inter alia*, AI may be an efficient cause but also may be its own material cause; categorisation of cause within a typology may become blurred.

[45] It could also have quite different quasi-neurological and hence sensory relations to the world, following patterns found in the animal kingdom, though not necessarily these. Some species of octopus (as cephalopods) have approximately 500 million neurons, but more of these are in their arms than 'brain' and experiment indicates some species have high capacity problem solving skills combined with a radically different sensory experience of the world (Godfrey-Smith, 2017)

organizations depend as much on our desires as they do on our capacity to override them (they become the root of aspiration and so are delayed or altered rather than repressed). Marketing would make no sense without desire, and yet marketing is a central component in real socio-economic systems. One might go further and note that aspects of the 'gap' as stated by Searle imply also the integration or perhaps more accurately interaction of the individual and aspects of society through deontology that depend on the relation between life projects of individuals and the aspects of society (however named) that exist as organized components. This is not to trivialize human projects as simply products of desire (by suggesting they lack reflexivity, durability etc), but rather to suggest desire may be something different or absent for an AI.

Desire in general is expressive in a way that having a goal only need not be. It is a bodily relation of thought (see Damasio, 1994). The body of an AI as a seat of thought could be quite different here and one might extend this thought about thought and the body to many other differences. Though the purpose of sleep remains a matter of dispute, recent research indicates one of the processes that occurs during sleep is β-amyloid clearance based on convective exchange of cerebrospinal fluid and interstitial fluid. Put another way, sleep (a potentially dangerous period of required inactivity) involves a shutdown where neurotoxic waste products produced whilst awake are cleaned from the central nervous system (Xie et al, 2013). However, this 'shutdown' is also a period of dream-states. One may reasonably ask on what basis an AI would shutdown and whether this is sleep from which dream-states could arise? In a human, dream-states are important for waking states as sources of inspiration, reflexive change and many other consequences. Would a self-conscious AI have a subconscious and an unconscious?

One could go on listing points of potential difference because of constitution. Alternatively, one could note that much of the above requires one to assume that an AI can in fact think, an assumption that takes for granted the technological capacity to create thought (Turing's aspiration completed), but simultaneously questions that the technology can be assumed to resolve other problems of (through duplication) the self (an embodied thinking feeling human-like-as-human-similar/emulating entity). As such, the focus on difference splits difference based on a non-necessary divide in what is assumed (a positive solution to one aspect, a negative for another). However, it is precisely because assumptions regarding duplication in one aspect do not require assumed duplication in the other that the potential differences can be explored. It is why they have been a staple of science fiction and of futurist speculation for decades.[46] It is also why such speculation has also been translated as critique of the Chinese room thought experiment (most obviously via the robot objection). In TH

[46] The speculation sits within a broader universe of philosophical-as-speculative argument. For example, the possibility that our reality is a cosmological virtual reality space: if a material species in a material universe survives long enough to achieve advanced technology it is reasonable to assume it will also produce computer technology, and this will involve exponential advances in simulation; at some point advanced computers could run indistinguishable 'real' simulations. This being so one might then assume the ratio of computer simulation realities to material realities favours computer realities, and this would imply that the likelihood is that any given species that is self-aware, such as ourselves, lives in such a 'reality'. As a further step, any self-aware entity that evolves within a virtual reality will develop science to interrogate that reality. The closer investigation comes to the fundamentals of that reality the more it will be revealed that the basis of the reality is mathematical (a strong Platonic claim for the status of mathematics inherent in a coded/designed/synthetic virtual reality).

these are the seductions of a world we are on the cusp of creating. None of which is to suggest explicit ontology has been the typical point of departure.

The immediate point, however, is that subcategories of AI[S] are also dividing lines. Searle's approach has been significant as a key point around which disagreement could coalesce, and has in this sense helped to create and foster the subcategories of AI[S]. Though Searle's Chinese room is helpful in setting out difference (what an AI does not demonstrate about a human) it is less helpful in terms of what may be the different constituent aspects of an AI once a dividing line has been drawn. It is under elaborated in this way (however, see Preston and Bishop, 2002). Posed in purely philosophical terms this may seem unimportant, since Searle's original argument regarding whether AI is in fact thinking has not simply disappeared. But the problem is not just philosophical, it is sociological regarding the consequences of philosophy. The very separation into subcategories has become a problem in this sociological context. Working across the subcategories has become a challenge, partly because there is something disempowering about the initial location of the majority of a subcategory within science fiction and futurism. One is constantly dealing with what can seem simultaneously profound, but also overblown and perhaps unserious. This is by no means to denigrate interesting work that is done. For example, in philosophy Shanahan (2010) on embodied AI, or Sparrow's (2004) alternative to Turing's game, which takes its inspiration from Blade Runner's Voight-Kampff test to create a thought experiment to assess the moral capacities of an artificial entity. But consider the broader problem of context based on increasing recognition of problems and potentials of the *actual technologies* that are developed under the aegis of AI.

Here one might note the 2017 23 Asilomar AI Principles that are intended to guide the future development of AI.[47] These heavily emphasize control, benefit, common good, risk assessment and caution. However, the principles are not legal injunctions, nor do they refer to intrinsic (or set binding extrinsic) limits to technology, nor can they prevent alternative interest-incentives that may subvert the principles (a state's concerns with security, surveillance and superiority of arms; a corporations concerns with competitive advantage and market capture).[48] It remains the case as we first noted early in the essay that the subject matters of second subcategory AI[S] are gradually coming into the purview of organizations, but doing so is a socio-political activity that is affected by the dominance of concerns that have accompanied the development of categorizations. Recall the case of EU deliberations on AI, the idea of an electronic person was not central, and yet was recognized to be increasingly important to address. Manifestly, based on the development of categorizations, the problems are multi-faceted: normalisation of AI, the issue of 'cope', functionalism and function, but now also dividing line inertia's that reduce the urgency or resist the centrality of concerns with real technological developments that may be occurring under the aegis of AI. One might want to consider this also in terms of other essays'

[47] https://futureoflife.org/ai-principles/

[48] For example, cyberwar and the new securitisation are deep problems. There is no mutually assured destruction approach to cyberwar, and this makes self-restraint logics difficult to implement. Cyberwar is not just information extraction putting security operatives at risk. It is also an ability to attack and paralyze electricity systems, hospitals, welfare administration or any complex bureaucracy. It is the capacity to manufacture fake news as interventions in democratic processes. These wars can be fought in proxy; dumping information onto Wiki-leaks… Cyberwar information is power but not truth.

comments on tendencies towards TH. In any case, it is not a matter of blame to note that Searle's Chinese room is limited as a resource in terms of any reasonable analysis of actual AI, but it is important to note that its discursive role has been important.

From Searle to real problems of 'AI'

Furthermore, once one starts thinking about sociological nuance one also starts to think about the way this inheres in social ontology, and this brings us back to matters of elaboration. For Searle, his social ontology is a philosophy *for* the social sciences not a philosophy *of* the social sciences; it is a simplified (rather than simplistic) apparatus that expresses the common constituents from which social reality is built. The claim is that all human institutional reality is created and maintained by a 'single logico-linguistic operation': status function declaration, and so has a 'common underling structure' (2010: p. 201). Two problems arise. First, the position requires that status function declaration be fully descriptive of the structure and, concomitantly, second, the position requires that other and further matters are not significant for both the structure of social reality and what occurs on in and through that structure, where this is deemed to be actual constituted social reality, since if they are significant then the structure itself is also not quite the structure-in-operation, which seems like a tension or incompleteness if not a contradiction in terms of what structure is vis-à-vis *creation and maintenance*.

One needs to be careful not to traduce Searle here. Searle's social ontology (like all of his work) is brilliant. However, it also has its limits and like all works, its points of pressure. Searle has configured his claim regarding social ontology to be internally consistent but in so doing he has preconfigured it to be potentially misleading. The claim is that human *institutional* reality is fully accounted for and so created and maintained by an instantiated variety of his social ontology. Since Searle is the one to define institutions through rule construction, and defines rule construction as a consequence of status function declaration, then it is by logical consistency that the claim acquires coherence. But the inference is that social *reality* is fully accounted for, rather than it is an internal system of rule creation and reproduction that is accounted for. Coherence becomes credence. Social ontology is a concern with social being, *Making the Social World* and *The Construction of Social Reality* are titles that convey the impression that more than linguistically stated (statable) rule systems are being accounted for, unless all that is significant are such rule systems.

Though Searle's approach is stripped down and elegant it can also read like an attempt to construct a code for how society operates, where the system in operation seems to be dependent on so much more than the code, and so the code is not fully expressive of what society is. This seems slightly ironic if one considers how the Chinese room thought experiment is directed (code is insufficient for comprehension) rather than how it is formulated. Searle's structure (his social ontology as status-function declaration) does not internalize error, ambiguity, conflict, contingency, multiplicity, materiality, the constitution of distinct parts and their interactions through a formal account of emergence, and it provides little sense of the different varieties of transactions with reality a being is engaged in (bodily, personally, socially etc), which may also extend to how a person who is not simply an expression of the

sum of social constructs negotiates and lives through social constructions (and this must be important if Searle's 'gap' -- free will -- is to be real, and then also instantiated).

If status function declarations are all that is required to create and maintain social reality then everything else becomes superfluous, an etc or details (and much of this for Searle is delegated as 'background').[49] The formal operative potential of directed language use becomes the overwhelmingly significant aspect of society. However, one might argue that everything else matters to what actually occurs and so what things become *in and through time*. If one is to warrant the claim that a social ontology can be the philosophy *for* social science it must also be a philosophy *of* society.[50] It must be sociologically operative. Searle is confident that his approach is. Status-function declaration as social ontology does more than merely confirm an internally consistent claim regarding the logico-linguistic statement of itself as theory. The claim is also that it has more than merely *some* purchase on social reality. For Searle, it is *the* building blocks of institutional reality, and also a basis for explanatory investigation. Here Searle argues that institutions ground institutional facts, and behavior operates with entities referenced to these through rationality in the form of 'propositional structures' that express reasons for acting (every actual deontic power has a why and because potential for investigation).

However, though Searle's social ontology allows one to ask and answer important questions about society, it is questionable that it can provide appropriately developed explanatory accounts of what has and is actually happening *in* society (it has a *some* rather than *the* relation to accounting for and explaining society). This is despite that Searle claims in reply to Lawson that '*Most important, the analysis has to really analyze*. Nothing must be left unexplained' (Searle, 2016: p. 402).[51] Arguably Searle's social ontology produces sociologically limited accounts of social reality: it is an institutional fact that *a* is President. He had rights *b* and duties *d* within institution *e* where x counted as *y* in *c*. It is by virtue of rationality h that person j did g, and this was accounted for by deontic power p under reasons for acting n. One could describe Donald Trump in terms of deontic powers, institutions, institutional facts, and rationalities through reasons for acting, but arguably doing so would not provide a satisfying explanation (of personhood, life projects, integration into existing possibilities of institutions in decay, changes *through time* based on interactions in the

[49] This is implicit in Lawson's critique of Searle and occurred to me whilst reading that critique; notably, 'the sort of totality it is… has a bearing on the sorts of positions and power relations that will be involved' (2016: p. 388).

[50] Consider: actual operative deontology is not abstract logic it is also the integration of institutional conditions with possible life projects and ultimate concerns. This is a sociological problem not reducible to constitutive rules or institutional facts along the lines typically stated by Searle. It is the feeling of and for a system. Searle's logic of reason may not be ill-founded but the framing of reason seems insufficient to capture the fullness of human lived experience as social reality. As abstraction it is impersonal and disembodied and this seems to abstract from what seem extremely important aspects of the human condition in order to make claims about what conditions human existence.

[51] Searle's response to Lawson also seems to trade on an ambiguity between only found in humans and the only thing important for the constitution of social reality: status-function declaration seems to be unique to human civilization therefore status-function declaration is all that is significant for human civilization as constructed. It is not clear that this should follow, and so it cannot be the actual force of argument as claim that can be used to refute by reply the statement of alternatives that seem to share constituents with animals.

context of disintegration, creating unintended consequences etc). Trump is at the least a status dysfunction. As sociology the account would (without augmentation) be exsanguinated, as psychology it would be silent.[52]

Perhaps this seems like something of a detour from the focus on focus in terms of AI^{W} and AI^{S}. However, recall that the point is to assess the consequences of how arguments have been constructed, positioned and pursued. We have already noted that the problem of AI is not just philosophical, it is sociological regarding the *consequences* of philosophy. One might now also add that the problem is sociological regarding the *form* of philosophy and that includes ontology. Searle's Chinese room has played a major role in developing the sub-categorization of AI^{S}. His own approach to the Chinese room situated in terms of his ontology and social ontology provide limited resources as ways to think about any realized AI (since its constitution is likely to be different). One might now add to that, though the point is contestable based on competing ontologies, the ontology and social ontology within which Searle's approach to his Chinese room argument is situated (a reversal of the above phrasing) provide limited resources for addressing the problem of any actual technological changes under the aegis of AI. This is a different point than that the existence of the subcategories of AI^{S} seems to have created discursive constraints on addressing the problem of any actual technological changes under the aegis of AI (where the primary concern is not how actual 'AI' develops to function in the world). The point here is that the social ontology provides limited resources for exploring society *as a system in operation*, and 'AI' is an important source of change and challenge within that system.

If one is to consider processes in time then one needs an ontology of process and time with a developed and consistent methodology. Archer's realist morphogenesis seems an obvious candidate (see Archer, 1995, 2012). However, if we are to consider specifics, then the concept of relational goods seems a useful way to bring this essay to a close, since it allows us to follow on from contemporary change in and around actual AI whilst returning to the problem of focus on function in a more immediate way.

Function, AI, substitution, delegation and Relational Goods

The sense that AI is coming and we must cope is now widespread. However, there are various attempts to position this as intrinsically beneficial, or liable to be so based on the already existing processes that exist by which AI is developing or through which its development can be managed. If we return to the 100 years Stanford project report we began from, I suggested this exhibited an AI^{W} position, whilst it simultaneously acknowledged the role of law etc in relation to this, which I suggested was also a matter that evoked AI^{S} issues of entities and shaped the way some of those issues were stated and made more or less significant. We are now in a position to say more about this. According to the Stanford report:

> The measure of success for AI applications is the value they create for human lives… Given the speed with which AI technologies are being realized… the

[52] As philosophy the problem of logico-construction as presupposition, where primary statements define and confine subsequent developments, also creates a problem of transposition if one wants to consider the whole as necessary rather than sufficient for explanation.

> Study Panel recommends that all layers of government acquire technical expertise in AI… Faced with the profound changes that AI technologies can produce, pressure for 'more' and 'tougher' regulation is probably inevitable. Misunderstandings about what AI is and is not could fuel opposition to technologies with the potential to benefit everyone. Inappropriate regulatory activity would be a tragic mistake. Poorly informed regulation that stifles innovation, or relocates it to other jurisdictions, would be counterproductive… In privacy regulation [we advocate], broad legal mandates coupled with tough transparency requirements and meaningful enforcement – rather than strict controls… This in turn supports the development of professional trade associations and standards committees that spread best practices… (Stone et al, 2016: p. 10) 'Policies should be evaluated as to whether they foster democratic values and equitable sharing of AIs benefits, or concentrate power and benefits in the hands of a fortunate few… [Thereafter, AI must be introduced] in ways that build trust and understanding, and respect human and civil rights.' (Stone, 2016: p. 11)

Now, consider the context in which the statements above are made and also what they and the report in general do not state. The page 11 quote sits awkwardly with the page 10, since the latter (taking privacy as an archetype) emphasises that one should resist regulation until well informed and suggests the best source of such information is the best practice that emerges from trade associations and standards committees. It leads to a dominance of self-regulation in market situations by powerful private parties to those situations. In general, this assumes that information and practice are already, or are developing along, lines that are *objectively-as-universally* beneficial and that this, furthermore, is either normatively beneficial through development and discussion by parties or is a situation of normative neutrality in relation to technology (since more and better information and best practice are intrinsic to processes and these are associated). But this then requires that the driving force of change and innovation within the world and under the authority of the requisite bodies is expressible in these beneficial ways and that no other considerations can also exist that subvert, shape or co-opt what occurs and under what circumstances. This is a sophisticated way to express acquiescence or lack of resistance or further scrutiny to an ineluctable process of change in relation to 'AI' (where AI is a travelling frontier of technological development rather than an entity for which thought etc has been decisively demonstrated). It reverses the meaning of caution to mean 'do not be hasty in impeding change' rather than 'consider carefully what the consequences of change may be before they actually manifest'. Moreover, it restricts the capacity to participate, and so be powerful, to those who are already powerful by virtue of position as control of information or resources. It empowers and authorises those who own rather than those who are subject to consequences. Concomitantly, it creates a barrier to broadening and democratising deliberation and participation in the process by which change occurs and through which change is shaped. *AI is what AI does* and *AI will be what AI researchers do* acquire more of a problematic set of connotations when considered in this way. AI^{w} is not without strength if one starts to think about power relations, and here one might also go back to and rethink the context of the Pew research I previously referred to. In the US at

least, the public feel uninformed and unable to effectively engage with change in relation to AI (and TH).

Then consider the nature of contemporary capitalist societies in which the set of injunctions against hasty injunction are made. The value to human lives is not the measure of success in capitalist processes, so stating this as the measure of success of AI applications is potentially adverse arrogation rather than justified extension. A camouflage of concerned language disguises a basic logic that requires that the specific interests of some become the engine by which the interests of all will manifest. Now consider how the two are supposedly integrated, aligned or mediated: it is not in relation to citizens as citizens only. It is citizens as consumers, citizens as workers, and then citizens as further recipients of services (state provided or otherwise in the context of welfare). Let us consider these in order.

In terms of dominant ideology, it is as consumers that the many exercise power through markets, and so it is through the dynamics of such behaviour that a lightly regulated and mainly self-regulated system of corporations is shaped. The wants and needs of the many are responded to by corporations, and so corporations are shaped by what the many want and need. Corporations are disciplined in this primitive democratic expression of individual power that becomes collective power through its effect on the profits of corporations. However, the value of human lives is not the focus or goal of this system, it is deemed to be an unintended consequence of the interactions of the system. Rather than a public deliberation on what the value to human lives is *and* how this is expressed in terms of the concepts of want and need, we are left to simply assume that processes will deliver what we really want or need, and this will be of value to human lives. This implicitly entails benign or benevolent capitalism where technological change including AI is more-or-less conducive to human progress. It also disguises the asymmetry of power that is heavily weighted towards corporations.

Corporations shape what we want or need and shape the markets in which individuals supposedly exercise democratised marketplace power. Profit drives corporations to capture markets and limit and pressurise choices. The effect of real AI here can be multiple. For example, in the absence of 'net neutrality' (a prohibition on service providers manipulating access to the range of sources of services to encourage some over others -- a hidden market advantage through constituting the market of choices) then AI can be used to channel access and activity to anything that requires internet connectivity. Another example is that in the absence of prohibition, AI as smart algorithms can produce opaque artificial stupidity that is difficult to contest because of the apparent objectivity of big data and quantified metric based decisions (affecting everything from credit access because of credit ratings, to who gets fired based on 'performance' measures). What both these examples illustrate is that corporations can control the infrastructure of contemporary life through AI in ways that preconfigure the social world of the consumer. Thereafter, specific AIs can become necessary to participation in society and so necessary to employability, or acceptance through social normativity (apps, smartphones and chatbots all have this potential). All of which suggests that an information and best practice approach to regulation, devolving to self-regulation, as technologies develop under the aegis of AI, cannot be assumed to be a decentred form that leads to outcomes that value human lives. There is rarely a simple situation where consumers can choose between infinite

options with no consequence to themselves but every consequence for the corporation. An information and best practice approach favours those who are already powerful by virtue of position through control of information or resources, has the potential for human harms, and subordinates any concerns regarding the value of human lives to the values of corporations.

As I noted at the beginning of this paper there are grave potential problems here in terms of AI[w] and the problem of focus on function as a way to decentre or marginalise the important issue of what is important to human concerns or flourishing. In addition to the context of consumption, AI has major ramifications for work. A discursive split is beginning to emerge between those who argue that an imminent AI (and robotic AI) revolution will be transformative and liberating and those who argue it will be devastating. The former is typically expressed as an intent to 'take the robot out of human work rather than to put the human out of work' (tedious, repetitive, and onerously physical labour will no longer be necessary for some sub-set of humans). The latter is typically stated as a 'this time is different argument' (AI will affect almost all parts of economies almost simultaneously, including previously non-replicable skilled middle and upper income jobs such as accounting, law and medicine, preventing a widespread response of transition to some other kind of work, since there will be insufficient scope for that work --- capitalist creative destruction is this time going to be destructively destructive). The problems hinge on the issue of substitution of AI for humans within a context where the economic system requires sufficient humans to be employed to earn the income that then becomes the source of consumption that maintains the corporations that use the labour (whether it be AI or human). The basic challenge is that individual corporations are required to treat with caution or resist technology that creates a 24 hour a day workforce that does not get sick, does not retire, does not strike, and does not earn wages or seek terms and conditions, since if they do not resist then the system of corporations is adversely affected collectively by the self-interest of every individual corporation (though there is nothing new about this tension, since Marx was able to point it out 150 years ago). The collective consequences of that self-interest is then socio-economic collateral damage -- the potential for widespread socio-economic disintegration and disruption with real human costs, unless solutions emerge or are designed (such as a universal AI tax on production of goods and services in conjunction with universal basic income for humans, or the monetisation of state spending systems -- a radical new approach to the institutions of fiscal policy via money creation). Again, there is a clear problem of how to address the value of human lives here based on processes that are already recognized, but are not sufficiently centred as matters of concern for the populations of societies that seem set to experience the consequences (they are currently matters of latent anxiety rather than front-and-centre urgent debate). So, in the case of citizens as consumers there seems to be an adverse assumption that it is by opting in and out of markets that most problems will be solved, and in the case of citizens as workers, the consequences for workers wait upon the capacity of corporations to recognize the collective problem of their individual activity where it is their individual activity that markets encourage (the bottom line).

Awareness in the world of a coming AI revolution in the form of substitution and work is growing. People are also increasingly aware that AI involves an issue of delegation. That is, the taking over of activity by AI on behalf of humans. This is the

realm of citizen welfare (the nurturing of self and others). The way people are becoming aware of this is fragmented. There is a recognition that the use of AI may free up time: an 'internet of things' can coordinate, anticipate and undertake tasks on our behalf through AI -- everything from adjusting central heating, to shaping through suggestion and pre-selection what we may want to be informed about, be interested in, and consume, to managing a calendar and maintaining contacts with associates, colleagues, friends and family.

Clearly, the possibilities here can be positioned as potential benefits to human lives (and perhaps also to the environment). However, much as in the previous cases, one cannot assume that potential is realised or that it is without the possibility of adverse consequences. Delegation creates a host of potential problems. If delegation is subordinated to efficiency then time is freed up from things we may have failed to appropriately value in order to create time to pursue imposed functions that may be harmful to the nurturing of a fully realised human. Things forgone may have had value in themselves: the craft of making or completing something, the development of self through the thinking through and doing of something on one's own behalf, the pleasure and meaningfulness in engagement with others, and so forth. Curtailment, convenience and quickness of end product based on delegation are not necessarily the same as 'better'. This is particularly so if the process of delegation infantilises the self and leads to the freeing of time that is then captured. One should not neglect the Tomorrow's World fallacy: for decades media have been predicting that labour and timesaving technologies would result in greatly increased leisure and reduced work because more can be done in less time. Though it may be that future AI transforms work through substitution, in the meantime, the observed tendency of technological changes has been the capacity to compel more hours of work, and for some, this has been based on the capacity to work from anywhere at anytime (in connective employment) or to be called in to work at anytime (in 'gig' economies). In this context the value of human life is subordinated to the meaning and practice imposed on efficiency. Benign or benevolent capitalism can no more be assumed here than it can in terms of the citizen as consumer, not least because the two increasingly overlap in consumption-based and financialised societies of debt-dependence.

In terms of citizen welfare and delegation there is also recognition that many societies confront a demographic problem in the form of an aging population combined with reduced birth rates and disaggregated patterns of living, creating a problem of care for the elderly; AI and smart accommodation based on an internet of things, combined with robotics, are now being considered likely solutions to this problem. Ostensibly, this more than any other area seems one in which potential benefits will manifest. However, it still shares with all the other areas set out an ultimate problem of being and doing. There is a basic challenge that needs to be addressed in terms of how being can be nurtured regarding what is and what is not done. This is an issue that a focus on function cannot resolve unless we also consider the nature of the human that the problem of function is to be resolved for.[53]

[53] And so the problem of what is a person and in what sense they flourish and suffer is centrally important; there is great scope for development here of a naturalistic ethics (if what a person is affects how a person flourishes). Searle, for example, considers this in terms of human rights in Making the Social World, Andrew Sayer provides a general account of embodied needy beings in *Why Things Matter to People* and Chris Smith provides a set of constituents of a person in *What is a Person?* -- an

As I have argued throughout and cumulatively, AI^{W} defers the problem but with powerful consequences in so far as it favours what already exists in terms of tendencies, interests and power. Following Turing and Searle, AI^{S} creates a whole host of issues that never quite bring together a *central concern* with technology as is and the human who is affected. *Inter alia* ontology is rendered interstitial, and yet the issues are quintessentially a matter of ontology and social ontology. In terms of this final matter of the realm of citizen welfare one insightful conceptual innovation in social ontology is Donati and Archer's relational goods. Relational goods provide an important way to think about how the human is nurtured in and through social relations. For Donati and Archer relational goods are goods created and enjoyed *through* relations, they involve some activity which is its own reward but that also creates collective social benefits. Such goods are diverse and are constituted as the quality of a relation that arises between people, such as trust, as well as the quality of experience of cooperation, coproduction or collaboration. Such relations can be intimately inter-subjective and informal or more associative and impersonal, but in all cases the goods are not interchangeable with material goods, and do not consist in the product of the activity (Donati and Archer, 2015: pp. 199-200 and 207). They are constituted and enjoyed through the activity. As such they require development and nurture and become the products of enduring relations.[54] They cannot simply be created by law or dictate. They cannot be captured or appropriated by any given party and cannot be commodified, bureaucratised or marketised without the relations themselves being subverted in ways that corrode the goods that are otherwise constituted. They are 'pro-social' in so far as they contribute to the integration of society, but they also do not fit readily into traditional categories of the public or private sphere, since the former is associated with administrative provision of goods by the state and the latter with the marketisation of goods by corporations, neither of which captures the sense of what relational goods are or provides unproblematic grounds for the constitution of relations from which they arise. However, according to Donati and Archer, relational goods 'correspond to fundamental human needs' (2015, p 215) and 'If these goods are ignored, dismissed or repressed, the entire social order is impoverished… with serious harm caused to people and the overall organization [of society]' (2015: p. 203).

The concept of relational goods can appear amorphous, but this seems a consequence of what the concept is intended to articulate, rather than a failure of clarity. We intuitively recognize that there is something common to the positive quality of experience of relations and there is nothing mysterious about this. What is mysterious is the way society can both recognise and yet fail to foster such relations (what else is instrumentality, alienation, ennui or even anomie, if not an anti-human

---

approach then critiqued by Archer and also Porpora, since there is no clear sense in terms of which Smith's set is the set of required constituents, rather than merely a list of possible characteristics. Lawson also provides a variety of naturalistic ethics.

[54] In general, Donati and Archer claim that relational goods require: 1) a personal and social identity of participants (they cannot be anonymous for each other) 2) non-instrumental motivation of each subject; the relation must involve more than achievement of some end 3) participants must acquire or be inspired by rule of reciprocity as a symbolic exchange 4) sharing: goods can only be produced and enjoyed together by those who participate 5) require elaboration over time; a single interaction is insufficient for the relations 6) reflexivity that operates relationally - sharing is also of the sense of what it is that is shared.

relational failure). For example, following my comments on AI[S] early in this paper the European Union Civil Law on Robotics both marginally recognizes (in two short statements) and yet defers any meaningful comment on the quality of relations as goods, specifically in the form of care: 'the 'soft impacts' on human dignity may be difficult to estimate, but will still need to be considered if and when robots replace human care,' (EP, 2016: p. 4) and 'human contact is one of the fundamental aspects of human care… replacing the human factor with robots could dehumanise caring practices,' (EP, 2016: p. 9). What is clear is that a concept of relational goods provides a potentially insightful way to examine the issue of delegation in terms of AI and citizen welfare. In so doing it returns us to ethics, a perennial issue for AI, but does so on a broader canvas:

> The proof that today's public ethics do not involve a common good in a relational sense is found in the case in which, for example, the problems of peace, development, the environment, and also of new forms of poverty are not confronted as problems of concrete human relations between co-present subjects but are simply treated as 'things' to eliminate by marginalizing violent persons, punishing those who do not succeed in competing, banning polluters, and helping the poor with measures that promote passivity, Problems are confronted by putting people where they cannot cause trouble. These are false solutions to problems because they are not inspired by the common good in that they leave aside completely the necessity of involving poor and marginalized people… In the arena of social policies, it is now very clear that these modalities for confronting distress, poverty, and social marginalization are completely unsatisfactory. Peace, development, a clean and safe environment, a decent life for everyone - these are all goods that correspond to the relational character of these objectives: that is to say they can only be achieved together; they are not the sum of individual preferences… Relational goods are the key for moving from the *welfare state* to the *welfare society*.' (Donati and Archer, 2015: p. 217)

Conclusion

AI is one issue among many and in the end must be conceived as one aspect of one world. Searle is surely correct that social science is an investigation into a single (if multiply produced and constructed and disputed) world (2016). This is a claim he shares with realist ontology and social ontology. However, as I have tried to establish in this essay the problem of AI has not come together in any clear singular sense. Sophisticated origins ('yesterday's') in philosophy have had consequences. Foci have developed expressing bifurcations and marginalisations, and creating interstitial issues. Little if anything has been resolved, and so function has dominated in ways that are significant for what occurs whilst dispute continues. A whole host of critically important issues based on actual technological potentials have arisen. Tomorrow continues to be affected by today without (people in) today having any clear collective idea how it (they) will produce tomorrow. And yet there is content to process in so far as some have a very clear idea of their agenda, and they in the main are not thinking as citizens or being invited to (or inviting others to) deliberate as citizens.

References

Anderson, D. (1987) 'Is the Chinese Room the real thing?' *Philosophy* 62(241): 389-393

Archer, M. (1995) *Realist Social Theory: The Morphogenetic Approach* Cambridge: Cambridge University Press

Archer, M. (2012) *The Reflexive Imperative in Late Modernity* Cambridge: Cambridge University Press

Badmington, N. editor (2000) *Posthumanism: Readers in Cultural Criticism* Basingstoke: Palgrave Macmillan

Bostrom, N. (2016) *Superintelligence: Paths, dangers, strategies* Oxford: Oxford University Press

Bostrom, N. and Ord, T. (2006) 'The reversal test: Eliminating status quo bias in applied ethics,' *Ethics* 116(4): 656-679

Braidotti, R. (2013) *The Posthuman* Cambridge: Polity Press

Brockman, J. editor, (2015) *What to think about machines that think* New York: Harper

Caliskan, A. Bryson, J. and Narayanan, A. (2017) 'Semantics derived automatically from language corpora contain human-like biases,' *Science* 356, April: 183-186

Calverley, D. (2007) 'Imagining a non-biological machine as a legal person,' *AI and Society* 22(4): 523-537

Cabrera, L. (2015) *Rethinking Human Enhancement* Basingstoke: Palgrave Macmillan

Chalmers, D. (1996) *The Conscious Mind: In search of a fundamental theory* Oxford: Oxford University Press

Churchland, P. and Churchland, P. (1990) 'Could a machine think?' *Scientific American* 262(1): 32-37

Clarke, S., Savulescu, J., Coady, C.,Giubilini, A. and Sanyal, S. editors (2016) *The Ethics of Human Enhancement: Understanding the Debate* Oxford: Oxford University Press

Collier, A. (2003) *In defence of objectivity and other essays* London Routledge

Crockett, L. (1994), *The Turing Test and the Frame Problem: AI's Mistaken Understanding of Intelligence*, Norwood New Jersey: Ablex Publishing Corporation

Damasio, A. (1994) *Descartes' error: Emotion reason and the human brain* New York: Putnam

Dennett, D. (2013) *Intuition pumps and other tools for thought* New York: Norton

Donati, P. and Archer, M. (2015) *The Relational Subject*. Cambridge: Cambridge University Press

Economist (2016) 'From not working to neural networking,' *The Economist* June 25th

Elder-Vass, D. (2012) *The Reality of Social Construction* Cambridge: Cambridge University Press

Ellen, M. 'Everybody loves a sad song, said the master of them all,' *The Times* 12th November 2016

European Parliament [EP] (2016) 'Draft report with recommendations to the Commission on Civil Law Rules on Robotics' European Parliament 2015/2013(INL)

Fodor, J. (1992) *A theory of content and other essays* Cambridge Ma: MIT Press

Fuller, S. (2011) *Humanity 2.0: What it means to be human, past present and future* Basingstoke: Palgrave Macmillan

Funk, C. Kennedy, B. Sciupac, E (2016) 'US public wary of biomedical technologies to 'enhance' human abilities,' *Pew Research Center,* July 26th available: http://www.pewinternet.org/2016/07/26/u-s-public-wary-of-biomedical-technologies-to-enhance-human-abilities/

Gidley, J. (2017) *The Future: A very short introduction* Oxford: Oxford University Press

Godfrey-Smith, P. (2017) *Other Minds: The octopus and the evolution of intelligent life* London: William Collins

Harari, Y. (2016) *Homo Deus: a brief history of tomorrow* London: Harvill Secker

Harnad, S. (1989) 'Minds, machines and Searle', *Journal of Experimental and Theoretical Artificial Intelligence* 1(1): 5-25

Herbrechter, S. (2013) *Posthumanism: A critical analysis* London: Bloomsbury

Hauser, L. (1997) 'Searle's Chinese Box: Debunking the Chinese Room argument', *Minds and Machines* 7(2): 199-226

Kurzweil, R. (2000) *The age of spiritual machines* London: Penguin

Lawson, T. (2016) 'Comparing conceptions of social ontology: Emergent social entities and/or institutional facts?' *Journal for the Theory of Social Behaviour* 46(4): 359-399

Lassegue, J. (1988), 'What Kind of Turing Test did Turing Have in Mind?', *Tekhnema* 3: 37–58.

Legg, S. and Hutter, M. (2007) 'A Collection of Definitions of Intelligence' Technical report, IDSIA 07-07 Available: https://arxiv.org/pdf/0706.3639v1.pdf

Mason, D. (2016) 'Human enhancement: The scientific and ethical dimensions of striving for perfection,' *Pew Research Center*, available: http://www.pewinternet.org/essay/human-enhancement-the-scientific-and-ethical-dimensions-of-striving-for-perfection/

Mays, W. (1952), 'Can Machines Think?' *Philosophy* 27 (101): 148–162.

McGilchrist, I. (2009) *The Master and his emissary* London: Yale University Press

Millican, P. and Clarke, A. editors. (1996) *Machines and Thought: The Legacy of Alan Turing*, *Volume 1* Oxford: Clarendon Press

Millar, P. (1973), 'On the Point of the Imitation Game', *Mind* 82 (328): 595–597.

Moor, J. (2000) 'Alan Turing 1912-1954,' *Minds and Machines* 10(4): 461

Morgan, J. (2013) 'A humanist narrative less than the sum of its parts,' *Metascience* 22(1): 111-113

Morgan, J. (2014) 'What is progress in realism? An issue illustrated using norm circles,' *Journal of Critical Realism* 13(2): 115-138

Morgan, J. (2016) 'Change and a changing world? Theorizing Morphogenic Society,' *Journal of Critical Realism* 15(3): 277-295

Nagel, T. (1979) 'What is it like to be a bat?' pp 165-180 in *Mortal Questions*, Cambridge: Canto/Cambridge University Press

Nordmann, A. (2007) 'If and then: A critique of speculative nano-ethics,' *NanoEthics* 1(1): 31-46

O'Connell, M. (2017) *To be a machine: Adventures among cyborgs, utopians, hackers. And the futurists solving the modest problem of death* New York: Doubleday

Pinker, S. (1998) *How the mind works* London: Penguin

Pinker, S. (2015) 'Thinking does not imply subjugating,' in Brockman, editor (2015), pp. 5-8

Pinsky, L. (1951), 'Do Machines Think About Machines Thinking?' *Mind* 60(239): 397–398.

Rainie, L. Hefferon, M. Sciupac, E. and Anderson, M. (2016) 'American voices on ways human enhancement could shape our future,' *Pew Research Centre,* July 26th Available: http://www.pewinternet.org/2016/07/26/american-voices-on-ways-human-enhancement-could-shape-our-future/

Regis, E. (1991) *Great Mambo Chicken and the Transhuman Condition* London: Basic Books

Rescher, N. (2011) *Reality and its Appearance* London: Continuum

Sandel, M. (2007) *The case against perfection: ethics in the age of genetic engineering* Cambridge Mass.: Harvard University Press

Saygin, A. Cicekli, L. and Akman, V. (2000) 'Turing Test: 50 years later,' *Minds and Machines* 10(4): 463-518

Searle, J. (1980) 'Minds, brains and programs,' *Behavioural and Brain Sciences* 3(3): 417-457

Searle, J. (1985) *Minds, Brains and Science* Cambridge, Mass: Harvard University Press

Searle, J. (2002) 'Twenty-one years in the Chinese room,' pp. 51-69 in Preston, J. and Bishop, M. (2002) *Views into the Chinese room: New essays on Searle and artificial intelligence* Oxford: Oxford University Press

Searle, J. (2010) *Making the Social World: The Structure of Human Civilization* Oxford: Oxford University Press

Searle, J. (2016) 'The limits of emergence: Reply to Lawson,' *Journal for the Theory of Social Behaviour* 46(4): 400-412

Shanahan, M (2010) *Embodiment and the inner life: Cognition and consciousness in the space of possible minds* Oxford: Oxford University Press

Sparrow, R. (2004) 'The Turing Triage Test,' *Ethics and Information Technology* 6(4): 203-214

Stephan, A. (2006) 'The dual role of emergence in the philosophy of mind and in cognitive science,' *Synthese* 151(3): 485-498

Stone, P., Chair, (2016) *Artificial Intelligence and Life in 2030: Report of the 2015 Study Panel for One Hundred Year Study on Artificial Intelligence (AI100)* Stanford University

Turing, A. (1950) 'Computing, Machinery and Intelligence', *Mind* 59(236): 433-460

Wilczek, F. (2015) 'Three observations on Artificial Intelligence,' in Brockam editor (2015), pp. 121-123

Wolfe, C. (2009) *What is Posthumanism?* Minneapolis: University of Minnesota Press

World Transhumanism Association (WTA) (2005) 'Artificial Intelligence and Transhumanism,' Available at: http://itp.uni-frankfurt.de/~gros/Mind2010/transhumanDeclaration.pdf

Xie, L. Kang, H. Xu, Q. Chen, M. Liao, Y. Thiyagarajan, M. and O'Donnell, J. (2013) 'Sleep drives metabolite clearance from the adult brain,' *Science* 342(6156): 373-377