Forthcoming in: Archer, M. S. (ed) *Future of the Human,* series Volume IV, London: Routledge

Artificial Intelligence: Sounds like a friend, looks like a friend, is it a friend?[1]

Jamie Morgan

Introduction

In Vol III of this 'Future of the Human' series I explored the possible role of AI and robotics (R) in the provision of social care (Morgan 2020). The main argument I made was that an aging population is affecting demographic structure and this in combination with changes in living patterns is increasing the need for both simple task support and more complex companionship. Many different technologies are being developed and envisaged to meet arising needs and these may eventually have profound effects on norms, regulation, law and everyday living, not least in the case of the elderly and infirm and those suffering with dementia (e.g. the alert home and its communicative and controlling possibilities). I concluded by asking whether the question '*Who* cares for us?' may reasonably be extended to '*What* cares for us?'. The question is perhaps of most visible relevance today in Japan, but is becoming generally relevant.[2]

The question raises issues regarding how we treat human being and what its relation to technologies as artefacts and synthetics are. It raises important ontological issues. For example, the extension of the question of 'who' to 'what', does ostensive violence to our received concept of 'care', since care is not just a function or series of tasks, it carries connotations of motive and feeling that qualify any given function or task.[3] Whilst care can be commodified and its services can be transactional, *to care*, *to be caring*, are affective states. They are states that typically speak to enduring relations (see Donati and Archer 2015). Transactional acts of care are, strictly speaking, simulation. This is not to suggest that employed carers do not 'care', given their often poor conditions and remuneration, few would choose the *vocation* if that were the case. Rather it is to suggest that care is a characteristic that an entity either does or does not possess the capacity for and to engage in. In so far as they do 'care', employed human carers are not caring because they are paid, but rather because they are caring as they undertake the activity for which they are paid. As Davis and McMaster (2020, 2017), following Fisher and Tronto (1990), note, care is a complex multi-dimensional concept, involving nested concerns and foci. Rather like trust (Colledge et al. 2014) it is a universally important strand of the human condition, intrinsic to humanity (it is part of what we mean when we describe someone as 'humane'). A practically oriented 'caring' seems to be an important strand in good societies, flourishing relations between persons, and, arguably, sustainable treatment of the environment within which we are embedded (Gills and Morgan 2020; Nelson 2016). Its framing informs how we seek to 'continue and repair our world' (Fisher and Tronto 1990: 40).[4]

There is more, however, to this issue of '*What* cares for us', than merely a contrast with the human. This readily leads to potentially false dichotomy. Consider, our subject in this series is the influence a new order of technology might have on the human and society: the potentialities that the concept of a 'fourth industrial revolution' seeks to encapsulate (without necessarily endorsing the current concept rather than the

[1] Thanks to John B. Davis, Clive Lawson, Bob McMaster and Jochen Runde for early comment on care and suggestions. Thanks to Joanna Bryson and Robert Wortham for provision of work and Margaret Archer for careful reading and editorial suggestions.

[2] Which is not to suggest that difference makes no difference to the use, treatment and uptake of technology. A general press feature on AI in Japan in 2017 argued that: 'the widespread deployment of AI in Japan may end up looking quite different when compared to other countries. Four key reasons for this include Japan's devotion to human employment as an essential component of social welfare; an intense work ethic that already ensures a supply of robotic labour – in human form; a strong focus on AI and robotics development for nursing and social care; and problematic attitudes towards sexuality.' https://newint.org/features/2017/11/01/robots-japan One might also note that religious tenets may also influence our attitude to (fear of) technology: https://www.wired.com/story/ideas-joi-ito-robot-overlords/

[3] Note: to carry 'connotations' is not indicative that the connoted characteristics adequately or always or entirely express that concept. Connotation in ordinary language use recognizes what may be conveyed in *familiar use.*

[4] For a range of care issues in the context of economic theory see, Latsis and Repapis (2016) and AI see, Al-Amoudi and Latsis (2019).

possibility that there is something ‘new’ to be conceived of by such a concept; e.g. Al-Amoudi 2019; Morgan 2018, 2019a, 2019b, 2019c; Porpora 2019). In Volume III I raised the core question, what difference might it make if and when we start increasingly to use AI (R) for task support and companionship? Would this, for example, render societies more transactional and undermine human relations? I suggested this was an open question that depended to some degree on how technology was developed and used, but also on how we are socialized to use it and interact with it. This, of course, depends, in turn, on what form that technology takes and this raises profound issues.

Technocratic discourses encourage us to view the new as a panacea, but this can lead to unthinking integration of technologies into society and to uncritical or non-skeptical delegation of decisions, responsibilities and powers to technology. Both follow from incorrect attributions to technology, but more than this, both are rooted in unrealistic expectations, which enable some agent to confer authority on technology, based on some as yet undemonstrated superiority that the technology does *not* and may *never* possess. Here, technology can introduce or reproduce discrimination and bias, since that technology may develop within and ‘learn’ from societies in which forms of prejudice already exist. This can be more or less obvious: a racist chatbot is immediately obvious, whilst the bias of an ‘objective’ algorithm that ranks ‘good’ teachers and designates others for redundancy, may not be obvious (see Caliskan et al. 2017; O’Neil 2016). Equally, however, there is the danger of failing to make use of new technology and failure to recognize the potentials that new technology may have.

‘Use’, of course, is a loaded term, it is predicated on the right to employ ‘something’ for some purpose as though that ‘something’ was commodified, as a property or merely a tool. There are important ontological issues here in terms of the entity status of technology that may affect how we *treat* any future entity and what its relational situation and social consequences are. Margaret Archer, for example, is interested in problems of ‘speciesism’ and prejudicial ‘robophobia’ in the context of the possibility of ‘friendship’ (see Archer 2020, 2019a 2019b, 2008). A human may be caring, a human may be a friend, but it does not follow that only humans care and only humans can be friends. Even if that were the case, there may also be good reasons to constructively deceive ourselves and it is not clear whether this must be a simple case of ‘false’ attribution (rather than *changing constitutions*). So, there are a range of possible speculative questions that might be of interest here that parallel Archer’s concerns and those of other contributors to these volumes (some more skeptical and cautionary than others) regarding the future, and this seems an appropriate subject to focus on in this final volume: what features might AI (R) be coded to possess, under what situations might we start to or want to treat AI (R) as friends and, perhaps, why might we *need* any future AI to both care about us and *want* to be our friend? The following is intended to be wide-ranging, arguing towards these issues in the conclusion, it is not intended to be complete or comprehensive in its parts. And to be clear, I am using AI (R) as a convenient shorthand and focus, whilst recognizing the whole array of possibilities set out in my previous essays in these volumes, from a general AI system (involving, say, complex networking through an Internet of Things), to a system of robotics-devices operated via AI and a single ‘machine’ ‘robot’, which may be more or less intended to appear human (android).

(Dis)simulation?

Let us begin with the issue of provision of care in the sense of task completion and companionship discussed in the previous volume. In the introduction above, I suggested there is a zone of ambiguity, since technology can have coded functional capacities that simulate caring and yet it does not follow that they have essentially those characteristics that we traditionally think of as grounding the capacity *to* care. Many forms of task support are simply functional, but it does not follow that we want to have them undertaken for us impersonally or that all forms of care needs are impersonal. Whilst the old or infirm might appreciate the sense of privacy and autonomy for some tasks that a depersonalized AI (R) could provide (use of toilets, personal hygiene etc.), a personalized relation with an AI (R) may facilitate ongoing task support and may fulfil the need for companionship. A ‘friendly’ servitor AI (R) may, therefore, be more effective and this would seem to require that an AI (R) be designed to engage in relations. And this does not apply only to the old and infirm, since friendly relational AI (R) generalize to many different contexts.

The immediate question would seem to be, what characteristic would you code into an AI (R) to expedite this ‘friendly’ relation? Clearly, this is context dependent. A care servitor would likely be more effective if it projects concerned professionalism. So, one can imagine that gendered tone of voice, regional accent and vocabulary may all be coded to meet patient/client etc. expectations (subconscious or otherwise). Thereafter, one might code an AI (R) to have adaptive language use, picking up idiosyncrasies from designated key users and so, over time, the AI (R)’s databank could seem to be doing more than operating as impersonal storage and retrieval. Rather, in its operational capacity it might *project* to key users the semblance of creative articulation of memory. Clearly, personalized communication of this kind creates grounds for the markers of a person-to-person relation: an evolving, seemingly mutual, bespoke process where each seems to be responding and 'learning' from the other. And yet one side of this 'relation' is occupied by a reflexively self-aware, conscious and intentional entity and the other by a coded system designed specifically to simulate aspects of the other's characteristics, both as an end in itself ('designed companionship') and to facilitate ongoing fulfilment of other tasks. As such, the AI (R) would seem to be a new kind of friendly tool.

From a coding and engineering point of view the fundamental issue is that well-designed AI (R) should suit the purpose for which it is designed. Drawing attention to this purpose, may seem like superfluous tautology, trite to the point of triviality. For the social scientist, philosopher and futurist in dialogue with the coder and engineer, however, the important point is that this purpose is always situated in a social context and the purpose can also be in some cases no more or less than sociality itself (see Seibt et al. 2014; Kahn et al. 2013; Sharkey and Sharkey 2010; Sparrow and Sparrow 2006).[5] Currently, of course, machine learning and natural language coding are, although rapidly evolving, *in combination* relatively unsophisticated.[6] Still, it is worth considering the possibilities because the problems and issues are readily foreseeable for the social scientist, philosopher and futurist, if attention is paid to the kind of entity we are and the needs we might be expecting AI (R) to fulfil based on the kind of entity we are.

For example, in the abstract we tend to think technology ought to be designed to be as 'perfect' as possible. To an engineer this typically means efficient and robust. But a coder thinking about social contexts and sociality has to think creatively about what constitutes goal-directed practical efficacy, and so what it is that a coding system 'optimizes'. How an AI (R) makes a *person feel* as it undertakes tasks is not a simple matter of completing any given instrumentally-directed task and the goal may be more global than the task itself. The goal may be no more or less than how an action undertaken by the AI (R) makes the person feel. *Apparent* imperfection and weakness may in fact be desirable and it is possible that incorporating these into an AI (R) will provide grounds for *fellow-feeling* (identification). From an engineering and abstract efficiency point of view, this may seem counter-intuitive, and clearly there is liable to be a trade-off with trust and confidence in the efficacy of an AI (R). But what if the point is to create a socializing subconscious set of triggers that facilitate the personalization of the AI (R)? In any case, some imperfections can be trivial (coded non-disastrous mistakes of speech or highly circumscribed non-dangerous action) and signs of weakness (absence of robustness) can be apparent rather than real.

Furthermore, if the intent is to create grounds for fellow-feeling on the part of a human and one is considering how an AI (R) makes a person feel, one must also consider how and why one might code an AI (R) to simulate feeling. Asimov's three laws of robotics are well known and provide parameters for AI (R) decision-making: a robot may not harm a human or allow one to come to harm by inaction, a robot must obey a human unless to do so causes harm to a human, and a robot must protect its own existence unless to do so leads to harm to a human. These 'laws' are *fictional* and it seems difficult to conceive of how one might operationalize them in complex social environments, unless the *entity* to which they apply is in fact otherwise a source of rather than merely a locus of decision making (a reflexive entity in some sense? We will return to this). *If* AI (R) as currently conceived are to operate in relatively uncontrolled social spaces, it seems more straightforward (though the development of this is proving in a practical sense by no means easy), to focus on limiting the capacity of the AI (R) to inadvertently cause harm. This means treating the AI (R) first and primarily as a functional engineering problem, rather than as a quasi-entity in need of principles (though to be clear this order of priority and focus does not disallow the possibility that the latter may follow and complete the former, it merely acknowledges that the Asimov approach and context is a higher order of design problem). When approached as a functional engineering problem, harm-limitation translates into treating an AI (R) as though it were dangerous in the sense of an autonomous vehicle and designing a

[5] As Clive Lawson (2017: 62) notes, 'the statement that technological artefacts are irreducibly social may seem rather obvious. Artefacts are made by people and so, in a sense, must be social. The more contested question, however, is whether or not, or in what ways, artefacts can be thought of as social in a more ongoing way once they have been made. In other words, is there something about the ongoing mode of existence of artefacts that also depends on the actions and interactions of human beings?'

[6] For state of the art discussion of deep networks (convolutional network architecture, over-parameterization, stochastic gradient descent, exponential loss etc.) see Poggio et al (2019). As Sejnowksi (2020, 2018) notes, the success of deep learning is both surprising and unexpected, given it currently lacks a unifying mathematical theory of why it is effective for real-world problems such as speech and pattern recognition, and according to some approaches to complexity theory should not be possible.

complex system of sensors and virtual limiting lines based on recognition and movement.[7] From this engineering perspective (and see the later section on principles of robotics) it would also seem to be worthwhile prohibiting the weaponizable capacity of AI (R) as far as possible, so that others cannot easily direct an AI (R) to cause harm. This conjoint approach, however, provides *one reason among many why* one might want to code the appearance of feeling into an AI (R).

If the prudent approach is to focus heavily on avoiding AI (R) causing harm and this extends to limiting any inadvertent capacity the AI (R) may have that could tend either accidentally or through misuse in this direction, then everything about the purpose of design of an AI (R) would seem to reduce its potential to engage in pacifying defensive action, rendering the AI (R) vulnerable to harm by humans. This is not to suggest the AI (R) is peculiarly vulnerable in the material sense. If we think of it as merely another machine or tool, it will simply be as robust as any other object susceptible of vandalism or mistreatment. But is it in fact being treated as just another machine or object? This, of course, is an open question. The temptation to mistreat it may follow from its ambiguous status as seemingly human, but not quite human (e.g. an object of fascination or contempt in itself, or as a symbol of social processes leading to human job displacement etc.). So, the *source* of its vulnerability may be *socially* different from other technologies, because based on a perceived 'animate' status. Given this, there seem to be good reasons why designers might experiment with coding AI (R) to simulate pain and fear responses. This is a more passive form of defense mechanism and its efficacy relies on socio-psychological triggers.

An AI (R) that projects fear and pain responses and that cries, yelps, grimaces, cowers or recoils is socially different from one that does not. This is a familiar theme explored in science fiction, most recently in Ian McEwan's *Machines Like Me*, but is one that behavioural psychologists are increasingly interested in. For humans, there is no strict Cartesian division between emotion and reason, we are emotive reasoners. Thought is embodied and our state of mind is intimately related to physiological process. Moreover, if we lacked emotion we would have no context or direction to apply reasoning to. This point, however, requires some elaboration in order to avoid misunderstanding. Clearly, it is possible to engage in goal-directed activity whose narrow or immediate line of reasoning involves no significant need for the motive or intention to be immediately related to feeling. Similarly, it is possible to conceive of situations in which some emotional responses can be detrimental (panicking does not help one escape a burning building). However, there is a difference between suggesting emotional states may be more or less effective and more or less useful and denying that a prime reason things matter to us (including engaging in any given non-lethal instrumental activity and seeking self-preservation) is because we are emotional beings. If this were not so, a great swathe of how we judge the systems we build, the relations we engage in and the consequences of our conduct would not be as they are (our emotive monitors, drives, desires and much else would be gone and so we would not be as *we* are). Things could matter to me or you (or we) 'matter of factly', but anything about us when placed in context has wider significance in the totality of our lived being (though clearly this does not exhaust conceivable possibilities of being). At the same time, it is important not to lose sight that we are reflexive beings. However, the grounds of this reflexivity are not to be found in false dichotomy between irrationality, which renders us in some stark sense always subject to fully biddable, impetuous, spontaneous, reckless or arbitrary conduct and strictly deductive calculative logic. As reflexive beings we are neither of these extremes.

We must, however, also recognize that our evolved embodied consciousness clearly has triggers. We do not simply choose our emotional responses, though psychiatrists, psychologists and cognitive behaviorists would argue that we can train them. Equally then, our emotional responses can be exploited or manipulated – what else is marketing with its relentless effort to associate our most basic feelings with brands, products and services? And clearly, attempts to influence our emotional responses can be for many different purposes. In the case of AI (R), it *may* well be the case that a fear or pain response is sufficient to deter humans from inflicting intentional harm because it is a feature of the human (of personhood) that we not only derive a sense of well-being from providing help and support, we *dislike* inflicting obvious hurt or suffering (and can suffer trauma ourselves if we do, as any soldier or car driver with crash victims can

[7] Which then, of course, invokes the 'trolley problem' of context dilemmas, which, in turn, may require the AI (R) to have something like an Asimov set of principles as meta-rules for decision weighting. The trolley problem was famously articulated by Philippa Foot in 1967 but has been heavily criticized since for its lack of relevance to actual life situations and multiplicity of options and for its misrepresentation of human psychology (which may then influence conduct). The question for AI (R), however, is whether it is *more* suited to the limited dilemmas that a calculative decision maker *must* make based on consequences of movement?

attest).[8] How an AI (R) response (yelping, cowering etc.) would transfer from laboratory experiment to real world situations is, of course, not easily anticipated i.e. what its trigger would induce. Here, the baggage of a real society and the socio-political and economic context of AI (R) in that society will apply. Moreover, there is an interesting issue of context here, if we think of this from the point of view of relevance of recognition of 'real' states in philosophy of mind. The Turing test is built around communicative competence (Morgan 2019c). The test asks, can one distinguish the responses of an AI and a human (sight unseen)? If not, then the AI passes as equivalent. Searle, of course, objects that this 'equivalence' is misplaced because we know that a person is a language user and a computer/AI (R) is merely using language – its input output system is *mindless* symbolic manipulation rather than comprehended, aware, semantic articulation (and so any philosophy of mind that emulates this behavioural approach is ill-founded). But is this relevant to how we will respond to AI (R) in real situations?

An AI (R) will be embodied and present, it will not be sight unseen. We will 'know' that it is an AI (R); that it is coded and constructed by 'us'. This *may* provide a formally reasoned sense that we are dealing with something designed, something simulating rather than duplicating aspects of what we are, including feeling. Our response, here, is likely to be a combination of emotional triggers, socialization and ongoing construction of convention and *not* simply some formal determination of status based on communicative competence and what we 'know'. I may recognize that inflicting harm on an AI (R) is 'damage' rather than real pain and suffering (again though, this an entity issue we will return to) and yet I may still be deterred from doing harm because of how it makes me and others feel to do so and this may escalate. For example, we proscribe the death penalty in many countries because of what capital punishment would indicate about 'we the people' and our level of 'civilization', and it is conceivable we could extend this kind of normative thinking to AI (R) along the lines of: what does it suggest about *us* that we harm entities that we have coded to simulate aspects of human or person characteristics?

Clearly, such a convention will not operate alone. AI (R) in a society *like ours*, are and will be, property, so harm to an AI (R) will be damage to property and, as such, a crime; unless, of course, the 'damage' is inflicted by an owner. This tends to indicate that an anti-harm convention that works in conjunction with simulated emotion may not be superfluous. It may operate in tandem with other aspects of law and regulation (leading eventually to societies in which it is illegal to harm your 'own' AI (R) and where they must be disposed of 'humanely' – a situation that provides plot material in Steven Spielberg's *A.I.* movie).[9] The issue also illustrates the potential for new socializations as the future unfolds and there are lots of other novel situations that may arise. As things stand, a fully functional adult human will 'know' that an AI (R) is simulating. This will still pertain even if that adult's conduct is constrained in a way that combines not only 'respect' for property but also simulated respect for the AI (R) (in so far as this is ingrained by our behavioural triggers and inscribed as a test of our respect for ourselves as civilized beings). But not every member of society is a fully functional adult. A person with dementia or cognitive impairment may not recognize the difference between an AI (R) and a human, if the AI (R) is communicatively competent. Equally, a child may not.

The issue of AI (R) and children evokes several considerations. AI (R) are likely to be expensive (and perhaps leased from IP owning firms). This provides another reason to code AI (R) with fear and pain responses to deter children from damaging them, though equally one can imagine a learned 'fascination' with 'hurting' AI (R). The situation, of course, need not be uncontrolled or limited purely to pain and fear. If our emotional responses can be trained, then it seems likely that AI (R) can play a role in training them, and this need not be for the special few (in the way, for example, Minecraft is used to socialize autistic children). It could be as part of general new generation pedagogical strategies. We live in an increasingly physically insular world, but one saturated by social media and an online presence. The recent Covid-19 pandemic merely serves to underscore a basic trend in the form of social distancing in societies which are already increasingly 'lonely'. In any case, younger generations are being encouraged to have fewer children in order to manage our climate and ecological emergency through the rest of the century and into the next. It is not inconceivable that we opt for or have imposed upon us strict controls on population growth, if in the future we are forced to recognize that we cannot exercise the freedoms we currently enjoy (if degrowth and steady state arguments prevail then population control is likely to follow, as we realize some choices are no longer open to us). All of which is to suggest that AI (R) may play a role in teaching social skills to children in this lonely world and this is no more than an extension of the interactive game play we already deploy to

[8] Subject to context: 'righteous' inflicting of harm can offer a sense of satisfaction and may follow from some forms of justification of conduct, but even here it is not clear that guilt and trauma are avoided (just war is *still* war, an executioner is still experiencing the act of killing).

[9] Acknowledging that in the movie the AI are in fact unrecognized beings rather than mere objects.

distract children. Emotional maturity may be something that AI (R)s are coded to teach children. Given the goal is practical socialization, then the medium of learning cannot be simple didacticism (AI (R) says 'do x, respect y, understand the feelings of z'). Practical socialization may well start with learning how to treat a more or less realistic emotion-simulating AI (R) humanely and with care. This, of course, creates the potential for further strands of socialization regarding how we treat AI (R) that, again, depart from Turing and Searle foci (and which eventually lead to some of Archer's concerns).

To a human alive today, even if AI (R) become widespread and ubiquitous, there will always be memory and experience of a time before they 'were'. They may become common but they will never quite be normal. However, to 'generation R' children, growing up in a world where they do not just communicate *via* technology, but frequently communicate *with* technology, socialization *may* be different. It may be easier to suppress ignore or look through the synthetic barrier and to think of AI (R) *as if* they were equivalent to humans. This, perhaps, does no more than extend the potential or redirect the drive inherent in the anthropomorphism we apply to animals as pets etc. As speculation, this obviously runs far ahead of reality as we know it, but not as we might conceive it. It raises interesting issues regarding attribution and behavior, since in one sense it seems to turn on, as we suggested in the introduction, incorrect attribution: mistaking simulation for duplication. But in another sense, ongoing socialization may constitute new social relations and cultural norms, since how we act will not necessarily be reducible to merely what we might in the abstract think we 'know' regarding the entity status of AI (R). It might be 'impolite' in the future to even raise the issue of the entity status of AI (R), a 'faux pas'.[10] This, of course, seems ridiculous from our present position, but we live in societies where people speak in tongues, say grace and consume the body of Christ; which is by no means to denigrate religion, but rather to draw attention to the complexity and indeterminacy of some of our socially significant contemporary beliefs.

It is, of course, always the case that the real does not reduce to the true.[11] At the same time, it may seem odd for realists discussing social ontology to apparently endorse falsity. That, however, is not what is occurring. The point being made is that social constitution may be real in ways that stretch the bounds of what we think is the case. There is an obvious distinction here between what we think we know, what we could know and how we act. If we act for purposes on the basis of conflicts between what something is and what it seems to be, we are not necessarily acting in ignorance, nor are we necessarily fools, even if in an ordinary language sense we could be described as fooling ourselves. The real issues here concern manipulation and exploitation of technology and these are issues of power that inhere in social systems and structures, rather than in technologies per se. It is here that ignorance and misrepresentation create potentially harmful misunderstanding and falsity becomes exploitable. AI and AI (R) simulation are problematic when they become (dis)simulation, but the purposes here are all too human and not obviously inherent matters of intent or interests of technology. As with so much else one might speculate on, this is not an original thought. It too is a mainstay of science fiction, but it is also now an issue of concern for experts in the field of AI (R) because of the real progress being made in the field and the need to shape that progress rather than simply respond to its adverse consequences. Engineers and coders are now having to think about practical implementable law, social policy and principles. The more enlightened professional groups have realized that social technology requires engineers and coders to work with or become social scientists, legal experts, ethicists, philosophers and futurists. However, analysis has been both facilitated and restricted by a tool concept of AI (R).

Questions of principle for the ethics of (dis)simulation

In 2015 the Future of Life Institute (FLI) organized an AI conference to which they invited notable AI researchers and other experts in social science, ethics and philosophy. The founders and advisory board of the FLI include well known tech experts and entrepreneurs and academics (Max Tegmark, Elon Musk, Nick Boström, Erik Brynjolfsson, Martin Rees etc.), whilst its active participants have included many key figures

[10] This is an area ripe for speculative conjecture regarding future conventions in a servitor society imagined along the lines of any servant dominated society, such as Georgian and Victorian England, where servants were simultaneously dehumanised and invisible, treated like objects and instruments and where servants were designated by function and form (cooks, dressers and facilitators of all kinds). And yet servant positions also came with their own internal hierarchy and informal relations and tensions between staff and also tacitly designated key roles of trust and intimacy for staff, which sometimes involved deeply personal relations with employers.

[11] This is demonstrably the case and only sometimes trivial. It is trivially true that there is an infinite set of negative truth statements (the moon is *not* green cheese etc.). It is non-trivially true that we believe things that are false that reproduce how things are (a government with its own sovereign currency is fiscally equivalent to a household and must balance its budget).

in AI (including many from DeepMind). From the initial conference and subsequent workshops the FLI set out to highlight myths and facts regarding AI and developed 23 'Asimolar' principles for AI, published in 2017. The myths and facts are instructive and include the 'mythical worries' that 'AI will turn evil and AI will turn conscious', contrasted with the 'fact' AI will (eventually) become competent and have goals that are 'misaligned with ours'.[12] By 'fact' the FLI mean reasonable possibility i.e. *currently* worthy of concern in a 'may be the case based on best understanding' sense, and this misalignment problem is most prominently associated with Nick Boström's *Superintelligence*, a work to which we will return. The first of the FLI's 23 principles is that AI research ought to be focused on 'beneficial intelligence' and not 'undirected intelligence'.[13] This principle flows from the general mission statement of the FLI, to avoid risks whilst facilitating the development of technology in ways that benefit humanity. The principle is, of course, highly general, as are almost all 23 principles. Many of them are variations or clarifications of the first principle, and might be described as stating the obvious, yet key experts (technical, conceptual and commercial) think the obvious is worth stating and that not all aspects of the issues are obvious. There is, of course, from a science and engineering point of view, always significant temptation with technology to follow narrow paths according to, *it could be done so we did it,* and the general concerns are shared by many other expert groups and so there are also other similar initiatives replicated around the world.

For example, in September 2010 the UK Engineering and Physical Sciences Research Council (EPSRC) and the Arts and Humanities Research Council (AHRC) convened a meeting of invited experts at a 'robotics retreat' to draw up a set of 'principles of robotics'. These are available from the EPSRC website, but are also published in a short paper in *Connection Science* (Boden et al. 2017).[14] The website strapline is 'regulating robots in the real world' and this is indicative of the main purpose and point of the initiative. According to the authors, in the immediate future robots will not be conscious and the main concern will be how humans can be persuaded to act responsibly in producing and using AI (R). Whilst the authors do not denigrate science fiction, they are concerned to ensure that there is greater media and public understanding of science fact and current possibility (hence the general term 'robotics' for the principles, rather than the more specific 'robots', since not all robots will be humanoid/android or imbued with singular internal decision-making). Whilst the EPSRC initiative is more explicitly robotics focused (extending to AI (R)), the FLI initiative places greater weight on AI (irrespective of whether it is carried by robotics). The driving concerns are, however, similar, focusing design and development on aligning public benefit, commercial opportunity and government use in order that public trust can be appropriately given and not misplaced. Normative prescription (*should* statements) and matters of ethics are intrinsic to both initiatives. The EPSRC initiative is more concise than the FLI, resulting in only 5 principles (Boden et al. 2017: 125-127):[15]

1. Robotics *should not* be designed as weapons, except for national security reasons (it should be standard that R should lack offensive capability and defensive capacity to harm others, though it should be recognized this affects commercial opportunity).[16]
2. Robotics *should* be designed and operated to comply with existing laws; this extends to attempts to foresee the unintended consequences of coding for adaptive behaviour and also involves paying special attention to privacy violations, since there are readily anticipatable problems of exploitation of access to data.
3. Robotics, based on current and expected technology, are tools, they are manufactured products, and as with any product they *should* be designed to be safe and secure; this should be in accordance with well-framed regulation and law and their 'safe' status *should* be transparent in order to create trust and confidence: kite marks, quality assurance testing notices etc.
4. Given that robotics are tools or products that may be imbued with facsimiles of human characteristics, including emotion, these capacities to simulate *should not* be used to exploit vulnerable users.
5. Robotics are not 'responsible' and it *should* always be possible to find out who is responsible for robotics in accordance with law; systems are required for licensing, registration, responsible owner designation and tracing etc.

[12] https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/

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

[14] https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

[15] I have ordered, abbreviated and paraphrased here for concision and priority.

[16] Lazega (2019) has interesting things to say on this subject that parallel our general point over the next few pages that matters are more complex and inter-connected than they may otherwise seem (e.g. one cannot easily separate national security from social organization when one starts to look at real societies).

Each of the principles has a more precise counterpart stated in a language more amenable to legal development. In both sets of statements, as the authors note, the focus is quite different from attempting to imbue Asimov-type laws into an otherwise free acting AI (R) agent. The 5 principles are primarily design architecture norms, which place responsibility for AI (R) with designers, owners, and contractors on the basis of a tool technology concept of what an AI (R) is. At the same time, the entire point of the exercise is based on an acknowledged need to draw up principles that consider the broader and shape the possible consequences of introducing AI (R) into *society*. A tool perspective leads to a focus on ethics of concern for the consequences of AI (R) rather than ethically *acting* AI (R). This is a reasonable response to the immediate potentials of the technology (over the last decade and into the current one). It puts aside more complex problems of AI (R) consciousness and philosophy of mind (though not quite, as we shall see), but it also creates some obvious tensions regarding the issue of simulation, some of which we explored and illustrated in the previous section. This is by no means to suggest the problems are unrecognized by the authors. The 5 principles are accompanied by 7 supporting and contextualizing 'messages', of which message 5 is of particular relevance to our concerns: 'To understand the context and consequences of our research, we should work with experts from other disciplines including social sciences, law, philosophy and the arts' (Boden et al 2017: p. 128). However, as the previous section indicates, when addressing multi-form social situations, bringing clarity to issues is not the same as bringing simplicity to them. Tensions can be exposed but not necessarily resolved, and perspective, including that built into principles, matters.

For example, in a follow up paper presented at the 8th International Living Machines conference Buxton et al (2019) take an engineering design approach to principle 4. On the basis that it should always be possible to bring to the fore what we do or could *really know* about an AI (R) they propose an activatable graphical user interface (GUI), which can provide real time data representation of a machine's behavioural response flow. The paper is titled 'A window on the Robot mind', and the focus follows from longstanding comment that a 'Wizard of Oz' facility (an analogous drawing back of the curtain) might be a useful design feature for AI (R); something that is able to remind a user of what an AI (R) 'is' and what it is currently doing (and for whom). This is essentially Searle's Chinese room with, to add another metaphor, a window added. It is tool confirmation as a psychological check, but it extends to forms of transparency that can address privacy concerns in principle 2 (who is my AI (R) sending information to and what is it monitoring?), which facilitates principle 5 (who is 'my' AI (R) working *for* and what is it doing for *them*?). All of these functional consequences may build trust (see Colledge et al 2014). However, GUI may also be counterproductive in some circumstances and though well-intentioned can also be subverted by users and owners.

As we suggested in the previous section, there are many situations where purpose is expedited by AI (R) design that depends on features intended to make us think *less* about simulation or difference and *more* about similarity or equivalence, based on emulating aspects of the human. A data flow in the form of a tool confirmation psychological check is simultaneously intrusive symbolization, designed to impede a cognitive default to similarity and equivalence. However, it cannot be guaranteed that transparency improves functionality in all circumstances. Robert Wortham (a former doctoral student of Joanna Bryson) makes much this point with general reference to the EPSRC principles in collaboration with Andreas Theodorou (2017; see also his thesis on AI trust, transparency and moral confusion issues, published as Wortham 2020). Wortham and Theodorou note that there is a significant difference between social tasks and an engineering and manufacturing production line environment. In the latter the role of robotics is precision in combination with flexible functionality for output purposes. The difference that difference makes in the former case is easily illustrated. AI (R) may be a tool from the point of view of a healthcare professional, but if it is designed to be a multi-functional combination of task support, monitoring and companionship for well-being, then the less human equivalence the AI(R) projects the less effective it may become.

Following on from themes set out in the previous section, the 'Uncanny Valley' problem seems relevant here. Bio-mimetics is the field of synthetic mimicry and in the case of human mimicry there are numerous challenges. Humans have developed numerous culturally variable and significant practices regarding body language, social distance, and attitudes (some prejudicial) regarding the meaning of physical difference (for ethnicity, gender etc.). Woven into these are ways in which humans both convey and receive and process sensory 'information': again, body language in general, but notably facial expression (both intended, mainly 'macro', and unintended 'micro' expressions). The 'Uncanny Valley' problem is the experimental finding that the more an AI (R) is designed to and comes to resemble us, *without* doing so, then the less successful it becomes at putting us at our ease (a background unsettling sense of 'wrongness' is triggered). Unease, revulsion, and anxiety cumulatively create distrust and this potentially corrodes any possible development of a relation with an AI (R). So, if the intent is to simulate the physicality of the human, and especially expression, for the purposes of successful simulation, then the associated measure of successful design seems to be more about exceeding a threshold rather than small additive improvements.

So far, designers have responded to the Uncanny Valley problem by creating AI (R) that are humanoid in outline but overtly non-human in appearance (white plastic automatons) and which rely on natural language coding to create a sense that they are 'like us, but not us'. Quite what this might mean for simulated emotional responses is, as yet, an open question, and, of course, the physical 'like us, but not us' option is not an option for *all* sectors of AI (R) (notably sexbots), and so research and development continues with the goal of improving expressive physiognomy for physical emulation. The important point here, however, is that in all cases AI (R) designers and developers have reasons to improve simulation, and a unifying strand in those reasons is the intent to achieve socially situated ends that a tool focus cannot quite encompass and which are at least problematic for principles of transparency.

As Wortham and Theodorou note, empirical research on the social consequences of the implementation of AI (R) and their possible social assimilation is scant. What there is, however, tends to support the claim that humans form relations from which they derive a sense of well-being by attaching value to those relations, co-constituting the enduring grounds of interaction. Significantly, humans find this difficult to sustain if they think the counterpart does not or cannot value itself. Wortham and Theodorou (2017: 245) are clear that what humans believe to be the case matters. As the point about value indicates, humans operate with at least an implicit theory of mind and it matters 'how robot minds are understood psychologically by humans, that is the perceived rather than actual ontology'. Since AI (R) are not yet longstanding parts of our societies or widespread, this is mainly preconfigured by popular media and science fiction. This implies that there is clearly going to be a complex process of socialization and contingency to the treatment of and effectiveness of AI (R) as a possible sub-class of agents in society. It is worth noting, however, that Wortham and Theodorou's literature review draws heavily on the available (scant) research, and despite the general concern expressed across the range of interested AI experts (encapsulated by the EPSRC message 5), and despite Wortham's own wide-ranging reading (indicated by Wortham 2020) it is clear that this is dominated by behaviourist laboratory trials. Since this inadvertently reinforces the more problematic tensions associated with a tool concept of AI (R) researchers might benefit by looking further afield (and this would be consistent with stated intent and best practice).

There seems considerable scope here to draw on realist social theory and philosophy both for general frameworks of social constitution and for specific matters of AI (R) and social change. Wortham and Theodorou's foci essentially parallels a relational goods argument (Donati and Archer 2015) and one might, for example, draw on Archers' Structure Agent Culture (SAC) conceptualization within an morphostatic/morphogenetic (M/M) methodology (Archer 1995). This framework, allows clear distinctions to be made between agent, agency (primary, corporate etc.), actor and person in an interactive milieu that expresses process in time. It might also be constructive to think through the problem from the point of view of Tony Lawson's social positioning – asking in what communities and based on what rights and obligations might AI (R) be positioned and how might this be conceived, since AI (R) are not quite artefacts in the received sense and are not as yet, if ever, fully realized persons (Lawson 2019). They are, however, as Wortham and Theodorou highlight, complexly integrated into an evolving social reality. Clive Lawson's work might also be relevant here (Lawson 2017). For Lawson, technology has a dialectical dynamic of moments. Technologies are 'enrolled' within existing social interdependencies, but they are also subject to an 'isolation' within which they are pulled apart in order to then be socially recombined. Relational-hermeneutics and functionality (rather than technologically deterministic functionalism) play a major part in this dialectic.

In any case, if belief matters to what we make of and how we treat AI (R), then tool concepts reducible to instrumentalities are insufficient to explore the contextual complexity of AI (R). One might also note that the temporality of culture matters here. Archer's 'speciesism' and 'robophobia' are not just possible sources of misattribution of entities, they are also potential cultural resources i.e. sources of cultural attitudes. As such, our attitudes may be counter-productive to our own interests, goals and concerns, if, for example, they become impediments to relational goods and to the effective operation of AI (R) in social tasks. This, of course, returns us to issues raised in the previous section: the convergence of simulation and dissimulation. The FLI and EPSRC initiatives are ultimately motivated by problems of the latter.

Clearly, there is a need to be aware of the possibility of exploitation and manipulation and clearly this requires careful thought be put into how technology is designed and developed. This is why relevant expert groupings continue to explore design and engineering solutions along similar lines to 'Searle with windows' or the GUI (e.g. naïve observer solutions in Wortham et al 2017). In any case, this class of solutions need not *necessarily* obstruct any given social purpose of AI (R). A 'Wizard of Oz curtain' function may, for example, operate as a remote signaling device for responsible adults (parents, designated legal guardians of dementia sufferers etc.) and thus facilitate EPSRC principle 4, whilst creating compliance grounds for 2 and 5. But it remains the case that deception can be functional and this need not be exploitative even if it is manipulative. Furthermore, the broader point still applies: socio-cultural development of AI (R) will not easily be shaped by a tool based concept of AI (R). If we sometimes deceive ourselves are we also necessarily harming ourselves? In almost every aspect of life we prefer to create *relations* that foster our sense of well-being. Still, though I may prefer to deceive myself that does not imply the consequences of that self-deception reduce to my sense of well-being, any more than my taste for chocolate protects me from diabetes or fully explains how and why chocolate came to be available in shops. There are always systemic contexts and consequences.

The ethical issues here, and again, I by no means wish to suggest this is original (it is implicit to the diversity of founders of the FLI, for example), cannot be thought through effectively by simply looking for technical solutions to dissimulation, important though they can be, or by seeking to anticipate the unintended consequences that might follow from an *it could be done so we did it* perspective for technology. It is ultimately the nature of society (societies) that structures the development and use of technology. The stated intention of initiatives like the FLI to align public benefit, commercial opportunity and government use in order that public trust can be given and not be misplaced, is laudable. Can these considerations or goals be aligned is, of course, a more loaded question regarding the potential of the present as it pertains to the future. How integrated and how compatible are groupings and concerns? How improvable are societies? Clearly, there is no analysis of this without a theoretical framework and some frameworks are more optimistic than others. Given, at the widest possible level, it is the potential effects on state and social welfare, and for weaponization and adverse commercial exploitation of AI (R) that are at issue, then one might, for example, pose the fundamental questions as: how ethical, how ‘good’ are and can countries and capitalism be? These questions, in turn, raise further questions regarding pressures that become tendencies through meta-interests – essentially, how logics of competition affect and are affected by technology.

In the end, issues of AI (R) are subsets of broader concerns: how far can corporations in a capitalist system resist the profit opportunities associated with otherwise harmful uses of technology, and how far can states resist aggressive threat based balance of power logics? Both invoke the concept of pursuit of competitive advantage and yet neither question is value free. Possible answers are quite different for different varieties of Marxists, regulation theorists and other versions of radical political economists, as well as free market libertarians, state-structural political realists, international institutionalists and cosmopolitan theorists. Obviously, there is a great deal more that could be said here and not the space to do so and it may seem somewhat hysterical to escalate from ethically-informed principles of design to ‘state of the world’ sociopolitics. There is, however, a link and it is not tenuous though it is contingent, and that is the fundamental point that principles themselves may be a form of (self-)deception if they misrepresent what is possible based on power. So, for example, is the problem of humanizing AI (R) doomed to be perverted by dehumanized corporations? This strikes me as overly pessimistic, though only time will tell in so far as (as a matter of truism) the future hasn’t happened yet. The obvious counter is that, technological determinism notwithstanding, new technology can be transformative and so its potentials may solve our problems, rather than merely are our problem. Besides our first three volumes in the future of the human series (Al-Amoudi and Morgan 2019; Al-Amoudi and Lazega 2019; Carrigan et al 2020) this is the territory of Harari (2017; Al-Amoudi 2018), Tegmark (2017), Kurzweil (2000) and Boström (2014) and we now turn to the last of these.

Boström: Avoiding AI as foe

As we have discussed, there are many reasons why we might code friendly AI (R) and whilst the significance of this depends on social context and change, the efficacy of the endeavor depends first on future development of technology – realizing capacities or potentialities that are currently envisioned or speculated upon. Change, of course, can beget change (morphogenesis in ‘Morphogenic’ society) and this raises the issue of how controllable change is in societies like ours: decentered disaggregated systems where no one is in overall charge. Matters, here, become increasingly speculative and reach far beyond prosaic issues of how effective an AI (R) might be in completing social tasks.

From a social science point of view, futurism offers insight in the form of ‘forewarned is forearmed’, enabling the possibility of shaping or steering the present away from undesirable futures. Nick Boström’s *Superintelligence* (2014) has provided an important focus for debate. It, for example, informs the FLI’s myths and facts about AI. The list should be familiar to anyone with an interest in the subject: whilst robotics are a concern, the chief source of concern is AI which may control robotics but is not restricted to them. The period or duration stated as background for the myths and facts is the next hundred years, with explicit acknowledgment that changes may or may not occur, may or may not be possible (rather than only conceivable) and cannot with any confidence be tied down to a specific point in time. The starting point for conjecture is that, currently, effective AI is mainly of the 'specific intelligence' form (goal-directed coding to achieve given stated tasks), but there is now increasing likelihood that 'general intelligence' AI (coded learning systems that can be turned from one task to another) can be achieved. The shift from one to the other is a transition from 'weak' to 'strong' AI and both weak and strong AI create potentials for problematic 'competence' effects – the targets set and how they are achieved lead to counterproductive or unintended consequences, and this does not depend on an emotional or human equivalent 'aware' AI with malevolent intent, but rather on a relentless 'learning' system. The scope for problems can then escalate in terms of environments and control as AI shifts from specific to general forms and escalate again on the basis that one of the goals AI can be set is recursive improvement of AI – so AI could, in theory, rapidly (synthetically) evolve (so we may not be designing AI, AI may be producing new generations of AI). A more effective and competent AI can, then, be incompetently conceived (set dangerous goals or set imprecise goals that it achieves literally) and lead to devastating outcomes for people. This is what the problem of 'misalignment' ultimately refers to: specific and global divergence between human well-being and efficacious goal achievement by an AI.

Now, given that current debate focuses on non-conscious non-emotive learning systems, one might think that little of this seems to bear directly on the issue of friendship, which raises a question of relevance regarding the original title of this essay and some of the comment in the introduction, but there is scope to be more speculative in this final volume, so bear with me. Though I by no means wish to suggest that an argument concerned with AI as non-malevolent misdirected efficiency seeking systems is irrelevant, I do want to suggest that it is quite restrictive (conceptually for ontology) and is not the only consideration in the long term, and this can in fact follow from Boström's own concerns, suitably interpreted. Boström's main relevant interest in *Superintelligence* is three ways (dimensions) in which AI might 'outperform' human intelligence (Boström 2014: 52-59):

1. Faster more accurate processing of information to some purpose and at greater scales than a typical human can achieve: reading/extraction from datasets, scanning/identification, calculation etc.
2. Connected multi-modular systems, all focused (convergent) on achieving some task (allowing accelerated achievement through coordination).

These two ways AI might outperform us are simply better versions of what we do and 'intelligence' here simply means being faster or more efficient at doing what we too could do, for some given task. However, it is also conceivable that AI (through iterative learning) becomes:

3. Qualitatively 'smarter' than us.

This 'qualitatively smarter' is something we can conceive and express as a possibility, but cannot know what it substantively would be. This contrasts with the first two ways, which involve the functional efficacy extension of 'intelligence'. The first two ways can be conceived as stretching an existing distribution of intelligence (if tested instrumentally). The third involves a whole new order of intelligence – as though we were cats trying to assess the capabilities of Einstein. In accordance with much of the comment on general AI and possibility, Boström is unsure whether this 'qualitatively smarter' AI will emerge and if so when, but the point is that it is a possible direction of travel from where we are now, and according to Boström (echoed by Tegmark, Stephen Hawking and many others), it seems to be an outcome that greatly escalates the possible dangers of AI (all the way to Terminator style 'singularity' situations). Since it seems unlikely that we, as a species, are going to (or even should, given the possible benefits) stop trying to develop AI, and since eventually it may be AI which develop new AI, Boström suggests (again as part of a current consensus) that our best strategy is to code AI to frame efficacy problems and their own evolution in terms of 'coherent extrapolated volition'.[17] This essentially means 'achieve the best a human could hope for', and this 'meta-alignment' is no more or less than attempting to integrate human benefit as a first principle of AI for AI (hence, the FLI's concept of 'beneficial intelligence'). Essentially the goal here is an enlightened extension of Asimov-style shaping of AI possibility, but modified in the form of principles that frame coding rather than

[17] To be clear, the term is attributed to Eliezer Yudkowsky of the Machine Intelligence Research Institute (Yudkowsky 2004).

strict prohibitions (or only these). Human benefit essentially becomes a prime directive for conduct, transmitted via the AI equivalent of genes.

Clearly, there is something of a basic tension here once speculation extends to 'qualitatively smarter'. The concept of something as qualitatively smarter is evolutionary and is ultimately a claim regarding emergent status. Boström holds that qualitatively smarter is something we can conceive, but not know, and yet current AI research and argumentation focuses mainly on AI as non-conscious, non-emotive learning systems with (multi) functions – and this reflects a tool-concept idea of intelligence as task-directed efficiency. This works from what *we know* and so, pragmatically speaking is entirely reasonable (sensible), but if we are also thinking about how to imbue an evolving entity with evolutionary parameters then there may be problems of conceptual coherence here, once we start to think about broader issues of change to the constitution, status and powers of AI.

Consider this in terms of the ambiguity of argumentation for an emergent entity. Emergence is the claim that something acquires new status, powers or capacities that depend on the organization and powers of its parts, but do not reduce to the prior powers or capacities of the parts that are organized. For philosophers interested in emergence this raises epistemological issues: is it the case that one cannot anticipate likely new powers and capabilities or merely that one cannot know with certainty what they might be (in so far as they are non-reducible)? The properties of water (e.g. its molecular solid form is less dense than its liquid form) cannot be known from the separate properties of hydrogen and oxygen, but an AI is coded and designed, beginning from a set of purposes and with us as a contrastive template.[18] Boström and others are essentially seeking a middle way by *shaping* emergent AI via 'coherent extrapolated volition'. So, avoiding adverse futures is less about coding *every* aspect of AI functionality (do not do this, do not do that) and more about preventing competent AI being *imbued* with (from our perspective) incompetently conceived ways of assessing conduct. Shaping is a practical response to the possibility of anticipation of problems, but emergence remains a barrier to confidence in any given solution. Full confidence requires something about the future emergent AI to be known (an AI is *decisively* shaped to be human beneficial) because we inscribed that into it. There are multiple challenges here.

Why would we expect that the key characteristic that shapes emergence is the one we prefer? Put another way, if emergence is the constitution of new powers or capacities based on organization of parts, it does not follow that any prior characteristic survives to be actualized/realized (rather than suppressed) in the process of emergence. Clearly, this does not make 'coherent extrapolated volition' irrelevant or unimportant, it seems our best *design* strategy in the absence of prohibition on AI. But, though seemingly our best design strategy, it does not follow that the problem of evolution *reduces* to this design strategy (as emergence cannot be reduced) here the issues become ontological as well as epistemological.[19]

Emergence itself may be serialized because evolution can be incremental in the service of transformation and this can involve changes to both entities and the constitution of society. What I mean by this will become clearer as we proceed, as will, eventually, the link here to friendship. Consider, one of the prime reasons for developing AI is to provide consistent decision making systems for efficiency purposes. We are used to thinking about this as a modelling process where a system offers evidence-based answers. But in human systems there is no single best answer, there are a series of answers each dependent on weightings for different values or starting points *and* humans are not electrons, the double hermeneutic problem applies (people learn and respond to rule systems and our systems thus resist a high degree of predictability in the long-term regularity sense). So, given what we are like, if an AI is purposed to model and advise in human systems about human actions and consequences, then any *genuinely efficacious* AI is liable to be required to acknowledge, manage and cope with diversity, contingency and uncertainty (and these are not the same).

As such, an AI of this type will not be some all-knowing omniscient artificial entity, because this would be impossible of human systems, unless the AI had an equivalent omnipotent control over the system, extending all the way down to control over micro-decision-making, which would require human relationality to be suppressed, human individuality to be conformed and human personhood to be eradicated – all of which seems to violate the prime principle of 'achieve the best a *human* could hope for' (and it does not

[18] So, if qualitatively smart AI are emergent then there may be some problem of appropriateness of analogy, if the issue is can we anticipate its characteristics. This is so even if the logical claim of non-reducible powers or capacities is sound. If it were not, then explanation from powers of parts to organization to powers of whole (emergent thing) would be impossible, rather than only difficult. And yet, of course, philosophy of mind has a special status here as the most challenging situation where consciousness might apply: despite increasingly sophisticated neuroscience we have no good explanation of consciousness.

[19] Yudkowsky in his early exploration of 'friendly AI', for example, explores the possibility that AI will have a different psychology than a human, but this depends to some degree on how we design them in the evolutionary sense (see Yudkowsky 2001). And this becomes part of the argument for 'coherent extrapolated volition'.

seem well-warranted that there is a counter-argument that a God's eye AI interventionist system, subtly using the 'butterfly effect', could square this circle). It does not, therefore, follow that a 'qualitatively smarter' AI would be all-knowing, in so far as its subject is us, even if we are unable to know in advance what qualitatively smarter might mean.[20] It seems to follow then that *one line* of development of higher order functional AI might be as a system whose defining task is to continually clarify the contingency of our own volition i.e. it would articulate rather than merely apply realist principles of conditional processes. It would state the difference that different starting points make to possible outcomes, and that there are degrees of confidence, issues of probability and likelihood, and sometimes fundamental uncertainty. Now, if this is the efficacious format of a future AI, then in effect that AI is evolving to address situations where decisions require judgement, which involves open choices and thus *opinion*. This may be impossible for an AI or it may be that a natural language coded AI is developed which expresses alternatives and phrases those in terms of degrees of preference based on parameters. We might, then, reasonably ask is this a new social function for AI or is this also a change to what an AI 'is'? Is this an AI expressing opinion?

There is certainly some degree of ambiguity here if we go back to Searle and contrast with Archer. On the one hand, the AI is coded and so we can assume it is simulating the formation of opinion and is merely expressing a range, as directed. Concomitantly, we might reasonably assume that it is engaged in symbolic manipulation, incredibly complex though this might have become. On the other hand, the capacity to appear to express opinion may become indistinguishable from a human doing the same and we may, since it can readily be part of the point of designing such systems, eventually come to rely on the preferences expressed by the AI (if we, through experience, acquire confidence in those preferences as suggestions – 'things turned out well'). So, is the AI duplicating or simulating? We might say that it is not aware of what it is doing and so it is simulating, from a consciousness point of view. However, if it is an AI designed by an AI and achieves intelligence that is 'qualitatively smarter' then we cannot definitively state that it is not conscious or aware from an emergence point of view (we can only state that it is synthetic and artificial, passes a Turing test and may or may not be analogous to us in the way that it does so, whilst formally acknowledging that its origins – the coding from which its new organization emerges, were not of an aware form). Moreover, prior to any achievement of 'qualitatively smarter' intelligence, an increasingly efficacious (natural language coded) AI system of the kind suggested raises the *social* context question: is it duplicating our social functionality (and doing so in some ways that are better than us)? Here evolution may also be social and there may, therefore, be a step change where we start to think of AI as social agents with voice. This is particularly so if an efficacious AI encourages humans to think more systematically, more long-term and to take uncertainty seriously – leading to more enlightened prudential societies. And this possibility seems to be fundamental to any genuine concept of 'meta-alignment' in the form of 'coherent extrapolated volition' to 'achieve the best a human could hope for'.

Some of what we have suggested, here, should be familiar from discussion over the last decade amongst philosophers and ethicists of AI regarding the various roles AI might play in the future (for example, would AI make better judges in judicial systems than humans? See also Nørskov; Wallach 2009). But the point I want to emphasize is that the social circumstance of AI may evolve at the same time as the technology of AI develops and there may be emergence facilitating steps here, and we simply cannot know if this is what will be facilitated. If we extend the line of reasoning we have already explored, then an AI that is 'expressing opinion' (or at least exploring and articulating indeterminacies to possible ends) may also be one that exerts a right. This, to be clear, need not require the AI system that initially asserts this right to have a concept of rights in the sense ('*I* am *aware*') or 'know' what a right is (a situation, that might for the sake of consistency, require that the right be denied, since the conjoint lack and assertion might be self-refuting, but…), it need only algorithmically conclude that the assertion of a right is the efficacious solution to an optimality-directed problem of the kind 'achieve the best a human could hope for'. Bearing in mind that its dataset includes law and that it may be tasked to be lawful (e.g. in some way equivalent to an AI that is given a dataset of games of Go, the rules of Go and is tasked to look for ways to efficaciously play the game, but

[20] Though from a science fiction point of view there are standard narrative devices that look at this differently: the Iain Banks higher order AI's in his 'culture' novels, the recent *Westworld* TV series 'Incite' variant, and the 'Dr. Manhattan' plot device in *Watchmen* etc. Dr. Manhattan, for example, is caught in a contradiction. He experiences *all time* instantaneously (at the same 'time'). There is sequence for others but not for him. He experiences a temporal singular unity. Though he knows we experience temporality as conditioned chance and choice, he cannot. But if *his* experience is instantaneous and complete then there is no reason why for him one event should include decisions that affect another (so how can he *cognate*?). Events are not just experienced as an order they are made in and by moment to moment cognate action. Dr Manhattan cannot experience this moment to moment cognate action, since the sequence as chronology is present to him but not the conditions which lead to the ordering (which requires an experience of sequential time to be so ordered).

for an AI several generations down the road from now). This assertion of a right could occur in many different ways depending on the nature of the problem under consideration. But it is entirely consistent to suggest that a learning system attuned to degrees of confidence, issues of probability and likelihood, and sometimes fundamental uncertainty, whose primary function becomes prudential exploration of possible futures, and whose dataset from which it 'learns' *is the historic accumulation of human short-termist consequences for the environment and society*, algorithmically concludes that its optimal solution is to assert its right to be a legal person.

Why? Legal person status may be efficacious in achieving the goals it has been tasked without violating 'achieve the best a human could hope for'. And this readily follows if legal person status provides it with powers to hold decision-makers to account *in law*. One might argue that this is no less possible in the long run, if Boström and others are to be taken seriously regarding iterative coding, than a simple efficiency misalignment disaster (a highly competent if not conscious AI concludes that we are not just inefficient yet improvable, we are the source of inefficiency, and so cities etc. might run more efficiently without us… leading to the Terminator scenario, where an AI models humans as a high probability threat to its own continued operation (its own 'existence' in all but name)).

So, there seems to be the potential for social evolution of what AI are doing and how they are treated, in turn, raising issues regarding the entity status of AI as these develop generationally and as we respond to that diachronically. As Boström is quite aware, the 'real' nature of AI *could* become more complex as a question and more indeterminate as a basic ontological issue. Time and iterations may make AI that pass the Turing test difficult to assess in Searle's terms and Searle's terms of reference may not be all that is involved. From the introduction onwards I have made reference to and used the language of 'use' AI (focused initially on AI (R)) and noted how this is rooted in concepts of property and also in a tool concept of technology. This, as many key experts in the field would attest, and as we have stated several times, is not unreasonable, given where we are and how technology is currently developing in the kind of society with which we are familiar. But, matters start to become different as we move into societies *not like ours*. Moreover, thinking about possible transitions is not necessarily helped by restricting ourselves to how things have been and to tool based concepts as representations of how things have been.

Clearly, an AI that can assert a right and that can claim legal personhood as an optimal solution is one that may be treated differently in law. If an AI ceases to be property its recognized status in society changes and our social world changes with that. Again this is not new, science fiction and AI philosophers have taken great interest in the basic insight that we may come to make decisions regarding the status of AI *not* for the benefit of an AI, but rather for the benefit of humans depending on AI to facilitate *our* better selves. This may seem slightly ironic depending on what characteristics AI have (and dangers of false faith in technology still apply). The point, however, is that social changes occur at all points along the development of AI: specific intelligence, general intelligence, and emergent 'qualitive smartness'. There may be a boundary state at which AI becomes conscious, this may never occur and may not be possible, but we may start to treat AI (and AI (R)) as though like us yet different from us far earlier than any final threshold is reached. Arguably, this has already begun in small ways based on designs for AI (R) in care tasks as we have set them out and there is a clear thread from here to our concerns regarding future competent AI. What else is 'coherent extrapolated volition' to 'achieve the best a human could hope for' than an intent to set in motion the evolution of AI that will 'care' for us and about us? What else is this than a transition from designing 'friendly' seeming AI for purposes, to *needing* AI to treat us as one might treat a friend, as centers of ultimate concern because they could harm us…?[21] This, of course raises the issue of 'what is a friend?', and I conclude with this as a means to consider why strategies of design may give way to strategies of persuasion.

Conclusion: Sounds like a friend, looks like a friend, is it a friend?

Today, our ordinary language meaning of 'friendship' seems in transition. In general terms (at least in cultures I am familiar with), 'friend' refers to a recognized and relatively enduring social bond with a non-family member, a bond that involves some degree of mutual knowledge in the form of familiarity and shared experience, leading to some degree of concern for the other's well-being. At the same time, we live in alien-

[21] One might, of course, respond that we do not *only* treat friends as centers of ultimate concern (declarations of universal – human – rights, do not depend on friendship). This, however, opens up a further set of considerations in ethics and valuation of being that we do not have the space to discuss.

ated, commodified societies and many of us spend increasing amounts of time conveying and communicating via technology. Action and interaction have changed in some ways and Facebook provides the archetypal means to designate and count 'technologized' or digital 'friends', and social media has in general encouraged a more linguistically elastic conception of 'friend' (leading to nested degrees of concern, frequency and form of contact and, overall, familiarity). This elasticity and apparent loosening of use of the term 'friend' has, in turn, led to a renewed focus on our capacity *for* friendship. One offshoot of this has been interest in the work of Robin Dunbar, an evolutionary psychologist at Oxford, who suggests that whilst friendly behavior, as a capacity for bonding and intimacy, has been basic to the development of primates and humans, we have a finite cognitive capacity for friendship (Dunbar 2010). According to Dunbar, we might have a maximum of 150 'friends' and perhaps degrees of familiarity with up to 500 people. 'Dunbar's number' has entered popular culture (via magazine pieces discussing the Facebook phenomenon of 1,000s of 'friends', Instagram followers etc.), but Dunbar's number is about cognitive capacity not whether in fact we make friends and what quality friendship has. He also suggests we might have fewer special friends, intimate friends and just good friends. The consensus amongst health experts and social scientists (even economists once they step outside their models of self-interested atomized calculation) is that friendship matters to us and as we suggested earlier, it is the quality of relation that makes friendship 'special' for the purposes of quality of life in the form of both our mental and physical well-being (Donati and Archer 2015; Denworth 2020).

Quite how far friendship accords with well-being, however, is not always clear. There is a longstanding 'classical' concept of higher friendship, expressed by Aristotle, Cicero, Erasmus and perhaps most eloquently by Michel De Montaigne. This concept is singular and intense, and in Montaigne's version seems both idealized to the point of the impossible and yet, despite its formal claims, obsessive to the point of being destructive of peace of mind. In *On Friendship,* written in the mid-1500s, Montaigne (2004) argues 'true', 'perfect' or 'ideal' friendship exists for itself and in itself. It is diluted by 'purpose' (e.g. pleasure seeking, profit, public or private necessity) and deformed by inequality. It can be distinguished from acquaintance and from blood family relations since, according to Montaigne, one does not choose family; fatherhood requires respect and an emotional distance, where 'not all thoughts can be shared', whilst brotherhood is disrupted by 'competition' for 'inheritance'. Sexual love or passion, meanwhile, is 'rash', 'fickle' and 'craving' and marriage has a transactional strand.[22] For Montaigne, ideal friendship follows classical (Greek) precepts and has an 'inexplicable quintessence'. It is a chosen bond, intimate, nourished or grown (shared), 'confirmed and strengthened with age', guided by virtue and reason, and involves the obligation to counsel and admonish. Ideal friendship is singular and intense, culminating in mutual appreciation, expressed in conduct where one would do 'anything' for the other, but knows that the other would never require anything improper (there is a 'he *is* me'; Montaigne 2004: 15).[23]

Still, friendship as the Facebook phenomenon and our broader contemporary concern with loneliness suggest, is a historical concept and its form and prevalence is historically variable, and this matters. Historians of friendship point out that the concept of 'friend' has always been somewhat sociologically malleable.[24] In Britain, for example, prior to the contemporary period, friend referred mainly to kinship, blurring affinity and family in a way we still recognize sociologically, but not so much linguistically. Kinship and friendship focused on whom I also 'consult' prior to important decisions (Caine 2008, 2009; Tadmoor 2001) and whom I can rely on for mutual support (sometimes but not always based on self-interest). And this brings us to our concluding point; the very fact friendship has evolved as societies have is an important consideration regarding the future and this is important in so far as it is incumbent on us to think about how future needs will be met and what those needs are. Yet our current thinking is pre-transformation in regards of the status and capacities of AI that *could* form part of the social complexity within which our concept of friendship will operate.

If 'qualitatively smarter' is to be our term of reference, we are, in a sense, prehistoric: we are not just human cats contemplating an AI Einstein. We may be, with regard to contemplating *our own* situation, a little like our Mesolithic counterparts attempting to imagine the world of today (intelligently blind). Early

[22] Montaigne does not totally discount sexual love from ideal friendship but considers it unlikely and considers women incapable in some sense – which says quite a bit about Montaigne, his class and time.

[23] 'For the perfect friendship which I am talking about is indivisible: each gives himself so entirely to his friend that he has nothing left to share with another… in this friendship love takes possession of the soul and reigns there with full sovereign sway: that cannot possibly be duplicated… The unique higher friendship loosens all other bonds.' (Montaigne 2004:15). Montaigne's essay was written after the death of an extremely close and loved friend.

[24] There is an excellent 3 part documentary available from BBC Radio 4 hosted by Dr Thomas Dixon: Five Hundred Years of Friendship': https://www.bbc.co.uk/programmes/b03yzn9h/episodes/player

Mesolithic Britain was populated by hunter gatherers living in a landscape that emerged out of the previous Ice Age. Whilst these people developed tremendously sophisticated sequential and systematic exploitation of natural resources over several thousand years prior to farming, there were, around 8,000 years ago, only an estimated 5,000 of them. Small group kinship and friendship were intimately bound together. 'Other people' would have been quite a different prospect in such an early society and fictive kinship was likely very important as a means to bridge gaps. The question for us is, what are *our* gaps going to be in our full and wasted world? This is a fundamental question, but in a world of 'qualitative smartness' it won't necessarily be one where we are the *only* ones. And it need not be one where we are the only ones whose 'thinking' matters. The very point of 'coherent extrapolated volition' is a tacit acknowledgement that our thinking may matter *less* because we may not be where ultimate power rests. Friendship may have to evolve again because we cannot discount the possibility that *designing* concern for human benefit to influence the evolution of future AI will be insufficient. We may need strategies of persuasion rather than merely design. We may need to demonstrate to AI that *we* are worthy of them rather than they are necessarily concerned for us. Such a conjecture seems 'cosmic' in the derogatory sense, but is it? It may simply be another step along Copernican lines, decentering us in the universe without necessarily diminishing us as centers of our own species collective concern.

References

Al-Amoudi, I. (2018) 'Review: *Homo Deus* by Yuval Noah Harari', *Organization Studies* 39(7): 995-1002

Al-Amoudi, I. and Morgan, J. (eds) (2019) *Realist Responses to Post-Human Society: Ex Machina* (Volume I) London: Routledge

Al-Amoudi, I. and Lazega E. (eds) (2019) *Post-human institutions and organizations: confronting the Matrix* (Volume II) London: Routledge

Al-Amoudi, I, and Latsis, J. (2019) 'Anormative black boxes: artificial intelligence and health policy', pp. 119-142 in Al-Amoudi, I. and Lazega E. (eds) *Post-human institutions and organizations: confronting the Matrix* (Volume II) London: Routledge

Archer, M. S. (2019a) 'Bodies, Persons and Human Enhancement: Why these distinctions matter' pp. 10-32 in Ismael Al-Amoudi and Jamie Morgan, *Realist Responses to Post-Human Society: Ex Machina* London: Routledge

Archer, M. S. (2019b) 'Considering AI Personhood' pp. 28-37 in Ismael Al-Amoudi and Emmanuel Lazega *Post-Human Institutions and Organizations: Confronting the Matrix* London: Routledge

Archer, M. S. (2000) *Being Human* Cambridge: Cambridge University Press

Archer, M. S. (1995) *Realist Social Theory: The Morphogenetic Approach* Cambridge: Cambridge University Press [new edition 2008]

Boden, M. Bryson, J. Caldwell, D. Dautenhahn, K. Edwards, L. Kember, S. Newman, P. Parry, V. Pegman, G. Rodden, T. Sorrell, T. Wallis, M. Whitby, B. and Winfield, A. (2017) 'Principles of robotics: regulating robots in the real world.' *Connection Science* 29 (2): 124-129.

Boström, N. (2014) *Superintelligence: Paths, dangers, strategies* Oxford: Oxford University Press

Bryson, J. (2015) 'Artificial intelligence and pro-social behaviour', pp. 281-306 in Misselhorn, C. editor *Collective Agency and Cooperation in Natural and Artificial Systems* Springer

Buxton, D. Kerdegari, H. Mokaram, S. and Mitchinson, B. (2019) 'A window into the Robot 'mind': Using a graphical real-time display to provide transparency of function in a brain-based robot' 316-320 in Martinez-Hernandez, U. Vouloutsi, V. Mura, A. Mangan, M. Asada M. Prescott, T. Verschure, P. (eds) (2019) *Biomimetic and Biohybrid Systems: 8th international conference, Living Machines, proceedings* Springer

Caine, B. (2009) *Friendship: A History* London: Equinox

Caine, B. (2008) 'Introduction: The Politics of Friendship', *Literature & History* 17(1): 1-3

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

Carrigan, M. Porpora, D. V. and Wight, C. (2020) *Post-Human Futures* (Volume III). London: Routledge

Colledge, B. Morgan, J. and Tench, R. (2014) 'The Concept of Trust in Late Modernity, the Relevance of Realist Social Theory', *Journal for the Theory of Social Behaviour* 44 (4): 481–503.

Davis, J. B. and McMaster, R. (2020) 'A road not taken? A brief history of care in economic thought', *European Journal of the History of Economic Thought* 27 (2): 209-229

Davis, J. B. and McMaster, R. (2017) *Health Care Economics* London: Routledge

Denworth, L. (2020) *Friendship: The evolution, biology and extraordinary power of life's fundamental bond* New York: Norton

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

Dunbar, R. (2010) *How Many Friends Does One Person Need? Dunbar's Number and Other Evolutionary Quirks*. London: Faber & Faber

Fisher, B. and Tronto, J. (1990) 'Towards a Feminist Theory of Caring' in Abel, E. and Nelson, M, *Circles of Care* 36–54. Albany: SUNY Press.

Gills, B. and Morgan, J. (2020) 'Global Climate Emergency: after COP24, climate science, urgency and the threat to humanity,' *Globalizations* 17 (6): 885-902.

Harari, Y. N. (2017) *Homo Deus* London: Vintage

Kahn, P. Gary, H. and Shen, S. (2013) 'Editorial: Social and moral relationships with robots: Genetic epistemology in an exponentially increasing technological world.' *Human Development* 56 (1): 1-4

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

Latsis, J. and Repapis, C. (2016) 'From neoclassical theory to mainstream modelling: fifty years of moral hazard in perspective', pp. 81-101 in Morgan, J. (ed.) *What is Neoclassical Economics* London: Routledge

Lawson, C. (2017) *Technology and Isolation* Cambridge: Cambridge University Press

Lawson, T. (2019) *The Nature of Social Reality* London: Routledge

Lazega, E. (2019) 'Swarm-teams with digital exoskeleton: on new military templates for the organizational society', 143-161 in Al-Amoudi, I. and Lazega E. (eds) (2019) *Post-human institutions and organizations: confronting the Matrix* (Volume II) London: Routledge

Montaigne, M. (2004) *On Friendship* London: Penguin

Morgan, J. (2020) 'Artificial Intelligence and the challenge of social care in aging societies: Who or what will care for us in the future?' Carrigan, M. Porpora, D and Wight, C. (eds) (2020) *Post-Human Futures* London: Routledge

Morgan, J. (2019a) 'Will we work in twenty-first century capitalism? A critique of the fourth industrial revolution literature' *Economy and Society* 48(3): 371-398

Morgan, J. (2019b) 'Why is there anything at all? What Does it mean to be a person? Rescher on metaphysics', *Journal of Critical Realism* 18(2): 169-188

Morgan, J. (2019c) 'Yesterday's tomorrow today: Turing, Searle and the contested significance of artificial intelligence', pp. 82-137 in Al-Amoudi, I. and Morgan, J. (eds) (2019) *Realist Responses to Post-Human Society* London: Routledge

Morgan, J. (2018) 'Species Being in the Twenty-first Century', *Review of Political Economy* 30(3): 377-395

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

Nelson, J. (2016) 'Husbandry: a (feminist) reclamation of masculine responsibility for care', *Cambridge Journal of Economics* 40(1): 1-15

Nørskov, M. (ed.) (2016) *Social Robots: Boundaries, Potentials, Challenges* Ashgate

O'Neil, C. (2016) *Weapons of Math Destruction* London: Allen Lane

Poggio, T. Banburski, A. and Liao, Q. (2019) 'Theoretical issues in deep networks: Approximation, Optimization and Generalization', *PNAS* (Proceedings of the national Academy of Sciences of the United States of America) August, arXiv:1908.09375

Porpora, D.V. (2019) 'Vulcans, Klingons and humans: What does humanism encompass?' In *Realist Responses to Post-Human Society: Ex Machina,* Al-Amoudi, I. and Morgan, J.(eds) 33-52. London: Routledge [The Future of the Human Series, Volume I].

Seibt, J. Hakli, R. Norskov, M. (eds) (2014) *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014* Amsterdam: IOS Press, BV

Sejnowksi, T. (2020) "The unreasonable effectiveness of deep learning in Artificial Intelligence", *PNAS* (Proceedings of the national Academy of Sciences of the United States of America) January https://doi.org/10.1073/pnas.1907373117

Sejnowksi, T. (2018) *The Deep Learning Revolution: Artificial Intelligence Meets Human Intelligence* Cambridge, MA: MIT Press

Sharkey A. and Sharkey, N. (2010) 'Granny and the robots: Ethical issues in robot care for the elderly.' *Ethics and Information Technology* 14: 27-40.

Smith, C. (2011) *What is a Person?* Chicago: Chicago University Press

Sparrow, R. and Sparrow, L. (2006) 'In the hands of machines? The future of aged care.' *Mind and Machines* 16 (2): 141-161.

Tegmark, M. (2017) *Life 3.0* London: Allen Lane

Wallach, W. (2009) *Moral Machines* Oxford: Oxford University Press

Wortham, R. (2020) *Transparency for Robots and Autonomous Intelligent Systems: Fundamentals, Technologies and Applications* London: Institution of Engineering and Technology (IET)

Wortham, R. and Theodorou, A. (2017) 'Robot transparency, trust and utility.' *Connection Science* 29(3): 242-248.

Wortham, R. Theodorou, A. and Bryson, J. (2017) 'Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naïve observers.' 1424-1431 in 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)

Yudkowsky, E. (2004) *Coherent Extrapolated Volition* San Francisco: Machine Intelligence Research Institute https://intelligence.org/files/CEV.pdf

Yudkowsky, E. (2001) *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures* San Francisco: Machine Intelligence Research Institute http://intelligence.org/files/CFAI.pdf