Friends, Robots, Citizens?

Stephen Rainey
Centre for Computing and Social Responsibility
De Montfort University
Gateway House, Leicester
+44 (0)116 207 7052
stephen.rainey@dmu.ac.uk

ABSTRACT
This paper asks whether and how an artefact, such as a robot, could be considered a citizen. In doing so, it approaches questions of political freedom and artefacts. Three key notions emerge in the discussion: discursivity, embodiment and recognition. Overall, discussion of robot citizenship raises technical, political and philosophical problems.

Whereas machine intelligence is hotly debated, machine citizenship is less so. However, much research and activity is underway that seeks to create robot companions, capable of meaningful and intimate relationships with humans. The EU flagship "Robot Companions for Citizens" project aims for "...an ecology of sentient machines that will help and assist humans in the broadest possible sense to support and sustain our welfare."[1]

This is a broad and ambitious aim, with a goal of making artefacts that can have genuine relationships with humans. This being so, in order to avoid merely creating highly interactive automata, the status of the robot must be carefully considered. Without significant public freedoms, for instance, the notion of a robot 'friend' would be a dubious one – as dubious as the notion of a 'willing slave', for instance. In a broad sense, these issues relate to the *politics* of robot kinship and sociality, perhaps specifically to civic epistemology. With a technological ideal of genuine human-artefactual kinship in the future, these political questions cannot be ignored. One approach to this problematic involves accounting for the robot citizen.

Categories and Subject Descriptors
K.4.0 General
K.4.1 Public Policy Issues, *Ethics, Regulation*
K.4.2 Social Issues, *Assistive technologies for persons with disabilities*
K.4.m Miscellaneous
K.5.0 General
K.5.2 Governmental Issues
K.5.m Miscellaneous

General Terms
Human Factors, Theory, Legal Aspects.

Keywords
Philosophy, Technology, Society

1. INTRODUCTION
The creation of a robot citizen cannot be achieved by purely technical means: citizenship throughout this paper concerns *taking an interest*. It is not clear that *any* amount of engineering can build this, nor have it be accepted as such by others. But there is an ineliminable technical part of the problem – not least in designing and building appropriate embodiment and cognitive faculties (it should be stressed that *cognition* is not purely facultative, though the faculties may be a necessary condition). However to truly achieve kinship with robots (as a shorthand for a variety of possibilities, e.g. machines, software, programmes), recognition must occur, or else every alleged companionship interaction would be dubious in the extreme. As with many aspects of human-human interaction (e.g. gender, race, occupation) the kind of human-robot recognition here required has a philosophico-political content that cannot be avoided.

This paper here hopes to develop a sketch of what would need to be the case for a robot to be considered a citizen, but not a manifesto and certainly not a guarantee.

The question of whether a robot *could* be a citizen is considered in terms of the conditions that would have to apply in order for a

[1] In the Strategic Partnership for Robotics in Europe Multi-Annual Roadmap (http://sparc-robotics.eu/about/) specific mention is made of "The ethical and social implications of social robots". In a broad conception of 'social', companionship and kinship between human and machine, human and programme, as well as inter-artefactual mutual reliances, partnerships, vulnerabilities and so on must be considered. Where genuine relationships are aimed at, discussions must go well beyond straightforward issues of human protection (http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5751970).

robot to share in place-making, where 'place-making' is an elaboration upon *merely* sharing space.[2] Citizens are explained as sharers of place, whereas *anything* can share a space with *anything* else.

The investigation will begin by looking at citizenship in very general terms, drawing upon Aristotle and Kant to substantiate the idea of 'taking an interest.' Drawing upon further philosophical thought from Searle and Habermas, the things in which citizens take an interest will be looked at. Finally, through the concept of embodiment, an exploration of *how* a robot could be thought of as taking an interest will be discussed.

2. CITIZENSHIP

The ability to contribute to the governance of one's political community is the notion central to citizenship in Aristotle's Politics Book III [2].

Aristotle makes a distinction between strict citizenship and qualified citizenship [2, pp1176ff]. The former can only be enjoyed by those free from service. This is to ensure that at any point citizens might be free to take part in governance. This is very much an active citizenship definition, one wherein the disposition toward political action is the marker of civic identity. Not all might enjoy the freedom to participate in governance that strict citizenship requires, notably in Aristotle's time slaves, women and foreigners. Foreigners could at least enjoy qualified citizenship. The point is none of these groups is thought of as being capable of contributing to governance, and so none can be politically active to the extent stipulated necessary for full civic identity

At the core, we can interpret beyond various ancient Athenian distinctions and say that citizenship is divided into at least two general groups which are citizens in a strict sense, and citizens in a qualified sense. For the citizen in a strict sense, the ability to take part in governance is a requirement. This in turn requires that those to be considered citizens must be free from impediments such as trades, poverty and service.

At least in principle, it would seem robots could easily fit the bill concerning freedom from trades, poverty and service. Were a robot to be constructed such that it had at least the semblance of free will, it would have no particular need to do any particular thing. That would rule out the need for a trade or service. Similarly, imagining a robot that was self-sufficient to the extent that many objects are, poverty would be no hindrance. It would not necessarily even be relevant. Yet, on this preliminary thumbnail sketch, this would not lead intuitively to an urge to partake in governance – where would be the impetus? This is one facet of the problematic which will be explored later, especially from section 4.

Aristotle's reasoning for granting unqualified citizenship to a particular group is that political society ought to exist for 'noble actions' and that these can issue only from a community, rather than from a mere alliance of various sorts of people. Aristotle's is a republican conception of citizenship, wherein participation or political agency is key. It assumes a fairly close agreement about ideas of the good life and about the various privileges of those involved in the community.

The republican model, in the shadow of Aristotle's Athenian ideal (Maybe typified by Florentine 'civic humanism'), may well be thought of as an impossible dream for modern, large, internally diverse and plural nation-states. If so, perhaps such republicanism can stand mainly as a critical standpoint from which to critique liberal political society. In fact, Kant can be read as hinting at something of a republico-liberal conception of the citizen, but on different grounds.

3. Kant

Kant suggests [9] that it is part of human nature that in society inevitable friction emerges as each individual seeks her own ends. This friction is offset by the claim that no single lifetime could feasibly accommodate the complete realisation of all of human beings' capacities. So, Kant supposes, the entire history of humanity is the arena wherein human beings' potential can be realised. This being so, politics is a necessary condition for human progress *per se* as it is politics that mediates the friction between the individual's plans and the progress of the community of all humanity. [5, p35]

In the context of an unfolding of humanity (of progress) and the necessity to act consistently with one's being an agent, one ought to do all one can to maximise the extent to which one can act and be unthwarted. From a historical point of view, social acting, on public reasons, is very important. Kant makes this point about law and freedom in terms of public and private reason, describing it as follows: Privately we must obey law, but always be ready publicly to challenge it:

> The public use of man's [sic] reason must always be free, and it alone can bring about enlightenment among men; the private use of reason may quite often be very narrowly restricted, however, without undue hindrance to the progress of enlightenment. But by the public use of one's own reason I mean that use which anyone may make of it as a man of learning addressing the entire reading public. What I term the private use of reason is that which a person may make of it in a particular civil post or office with which he is entrusted. [10]

At its most general, the importance of careful reasoning in terms of the public draws upon Kant's view on 'sensus communis'. This isn't 'common sense' as it would be known most widely, but rather is an *a priori* faculty of reasoning the denial of which would amount to a contradiction of agency in any given reasoner:

> ...under the sensus communis we must include the Idea of a communal sense, i.e. of a faculty of judgement, which in its reflection takes account (a priori) of the mode of representation of all other men [sic] in thought; in order as it were to compare its judgement with the collective Reason of humanity, and thus to escape the illusion arising from the private conditions that could be so easily taken for objective, which would injuriously affect the judgement. This is done by comparing our judgement with the possible rather than the actual judgements of others, and by putting ourselves in the place of any other man, by abstracting from the

[2] The focus here isn't on robot rights, a short history of which can be found at http://www.roboethics.org/icra2005/veruggio.pdf (bullet-points at the end point to the sources of concern)

> limitations which contingently attach to our own judgement. [11, §40]

The *sensus communis* is a form of individual judgement that takes into account others' partial ways of representing matters. The point of this is to scrutinise *particular* judgements in this light of *general* human reason. This should avoid the individual, partial perspectives on matters that although personally compelling, could in general have a detrimental effect on any judgment made. From this constitutive principle of judgements in agents *per se* springs a motivation for taking an interest in public (or social) matters, taking care in that interest, and hoping others do likewise.

The republican elements of this, similar to Aristotle's view, are that shared community is important and an interest ought to be taken in it. It is to be cultivated privately by respectfully following the law that structures it, and publicly by challenging those tenets of the law inadequate to it. The liberal part of this conception has it that politics, *via* law, protects the citizen in terms of their personal freedoms. Citizenship sees its domain in the interaction between the private and public spheres. Citizenship here is to be thought of distributively, rather than as an aggregative notion.[3]

In short, drawing on Aristotle and Kantian thought, citizens must be a community, with a sense of community, and at least be disposed to taking an interest in the governance of that community. This will be referred to as 'place-making', over above mere sharing of space. The persistence of place-making comes through the fact that many citizens are interested in how things are run, how they could be better run, and the ends at which the running aims. From this starting point, place-making can now be elaborated upon, and its conditions laid out in order to determine what would need to be the case for a robot to take part in it, thereby grounding the chance of citizenship.

4. Place-making

The simple-looking question, "Where are you?" offers at least two potentially controversial interpretations. On the one hand, the question can be answered in terms of space. Answering in this way might involve the reporting of a set of co-ordinates relative to a grid. The question might also be answered in terms of place. This could involve the reporting of a more varied set of factors. These can be seen in an example from 'Mediterranean studies':

> If the classic work of Fernand Braudel (1949) tends to view the Mediterranean over the long term as a grand space or spatial crossroads in exchange, trade, diffusion and connectivity between a set of grand source areas to the south, north and east, the recent revisionist account of Peregrine Horden and Nicholas Purcell (2000) views the Mediterranean region as a congeries of micro-ecologies or places separated by distinctive agricultural and social practices in which connectivity and mobility within the region is more a response to the management of environmental and social risks than the simple outcome of extra-regional initiatives. [1]

The first view is one that thinks of space in geometrical terms, whereas the second has a more holisitic view, drawing upon dynamic interests including the social. The latter is the place-making notion here put forward. An illustration might be helpful.

If Alice states that she is in Ireland, as opposed to in England, she means more than being simply further west than her London-based colleagues. Irish laws are different. Different customs operate. Different expectations mount when exploring Dublin as opposed to Dulwich, Dover (or Dresden, Darwin or anywhere else). This kind of difference between places over and above spaces is related to a holisitic notion of institutional reality, and social ontology, of the locations. This reality and ontology are the objects of interest for the engaged citizen.

4.1 Institutional Reality and Social Ontology

Institutional reality is a background to action for citizens. It offers a mode in which reasons can come which can warrant action in public. Reasons are required for action *qua* action and institutional reality provides a scheme from which are derivable desire-transcendent reasons (reasons not necessarily based in a metaphorical inner marketplace of competing personal desires) [14, p167ff] and a scheme *via* which desire can be articulated. Thus it presents a scheme from which action can result.

Searle describes how human beings have,

> ...the capacity to impose functions on objects and people where the objects and the people cannot perform the functions solely in virtue of their physical structure. The performance of the function requires that there be a collectively recognized status that the person or object has, and it is only in virtue of that status that the person or object can perform the function in question. [13, p7]

Status functions are concerned with the rules that constitute one thing as another in a context, e.g. a piece of paper as a payment in a shop, a person as a general in a war, a thrown towel as a submission. These things are *declared into existence* and so the normative power of speech acts makes something in the world that previously there was not. Searle here is flagging the *social* commitments and entitlements that constitute institutional reality. These commitments and entitlements serve to populate social ontology, moreover, as they are what create money rather than mere slips of paper; no-parking zones rather than mere tracts of land; a round of beers rather than a mere collection of liquids.

The collectivity of this is important. There is reciprocity at work without which these commitments and entitlements would become empty or disintegrate. There is then the question of who, or *what*, is taken as capable of ascribing, recognising and taking on social commitments of the sort that they can take part in the institutions of money, gambling etc. The question is actually familiar. Some people at some points are considered too young to take part in various institutions (in the UK, full time work for under 16s, 18 for gambling or drinking alcohol, 65 for receiving a bus pass). The roles are as much ascribed as the institutions: a bartender can sell beers she doesn't own, as she occupies a role given to her by the licensee, but she can't trade stocks in that role

[3] We can take from one of Kant's successors, Fichte, a parallel with his moral thinking, specifically the categorical imperative, and see the application for citizenship. The *I* and the *not-I* enjoy a mutual dependence, which is one reason for the necessity of treating others as ends on themselves – not to do so is to deny the mutuality of the *I* and *not-I* hence to deny oneself in a fundamental sense. Put briefly, to act against the other is not even to act but to be determined by a lack of understanding of what it is to be an actor, and agent, at all.

– there isn't a way in which my purchasing a Guinness can prompt Lloyds to remunerate the bartender or my portfolio to diversify. She isn't taken as a stock broker nor can she be both bartending for the bar and trading for Lloyds at the same time: illegitimacy in one or both roles creeps in. We can say from this that in occupying various roles and fulfilling them well, people mutually *enact* institutional and social reality.

The link between the enacting of institutional and social reality and place-making is worth highlighting at this point. It seems a clear advance beyond 'mere' space-sharing to be a constituent part of the possibilities for choice and action for oneself and others. In terms of the communitarian aspects of Aristotle and Kant mentioned above, this can also be seen in terms of taking an interest in the life of the community in that occupying these roles helps to constitute what can be expected in that very community.

4.1.1 The robot enacting social reality?

Whereas above it was asked from where the impetus to get involved in governance might arise for a robot, here the question becomes more complex:

- Could the robot be taken for something capable of ascribing status functions on one thing or another?
- To what extent could the robot *enact* institutional and social reality?

At the risk of littering the argument with too many rhetorical questions, this latest pair will remain hanging until the notion of *enacting* institutional and social reality is explored a little further. This can be done *via* Searle's notion of 'the Background' and the idea of 'civic nous'.

4.2 'The Background'

The role of social commitment as a structure to public action is ineliminable. There is a basic, possibly tacit, civic nous that guides interaction that can vary between place and place. In Searle, this is called 'the Background' [13, p135]. The Background is a set of mental states, not necessarily present to the mind at a given time, that sets up what the meaning of intentional states can be: they provide *the expected* from which divergence is noticeable (e.g. picking up an apparently heavy suitcase only to find it is a helium balloon shaped like a case. The surprise is analogous for the presence of the Background, though at any particular stage it wasn't called upon.)

Another way of thinking about it is as the set of justifications one would offer were one's routines to be interrogated (e.g. Why did you mime writing? I caught the waiter's eye – the bill has to be paid).

In terms of civic nous, the Background includes the capacity to recognise that standing on the left of a London Underground escalator constitutes a *faux pas*. The Background will underwrite spotting the error of an Englishman in Belfast inviting someone for a drink and actually having just one drink (rather than at least two). Whilst these sorts of examples might seem to indicate that the Background is simply a set of propositions, norms to be borne in mind, Searle argues [12, p156-7] that it is in fact not based in a mind-independent reality, but rather helps to structure the very reality that is inhabited. By way of another analogy it might be said that the flow of the Thames and that of the Soar look fundamentally alike. Invisible to the unaided eye is the bed that shapes and helps determine the unique way each river flows. In this analogy, 'the Background' is the riverbed, invisibly structuring the surface flow. It is something like a transcendental condition for the surface phenomenon – that *x* without which *y* could not be *y* at all – or of the Heideggerian 'immer schon' category that which is *always already* present.[4]

The London commuter doesn't hold in their mind any rules about escalators in order to avoid icy stares, having absorbed the fact that standing should only be on the right. The Belfast socialite doesn't consciously repeat the mantra that 'a drink' includes buying one back. Rather, each "...evolve a set of dispositions that are sensitive to the rule structure..." [13, p.145], where 'the rule structure' is the sets of social commitments collectively undertaken (without ceremony) in the context of the institutional reality in question.

Civic nous of this sort is the capacity to recognise contingency when the expected reality is deformed. Or again we might say that the Background manifests in social situations as that which gives content to surprise. This would suggest that contextualised cognition, this civic nous, is *embedded* in the institutional and social environment in which it appears, to allude once more to Heidegger, it is *always already* part of the logic of public being and public action, on which some more needs to be said.

4.3 Action

If we imagine a purely physical space of action, we can think of the laws of physics as the conditions for actions. The limitations of the body in contact with other surfaces are the limits of possibility here. The actions of the hypothetical dweller within the merely physical space are simply the instrumental interventions on the transcendent space. A physical space, if it is to be appreciated *as a place* in the sense here being used, will be a sphere of reasons besides.

Civic nous, the Background, collective status functions all come into play in a place. The intentions of the place-dweller, moreover, are structured according to the logic of the place: for instance, in wanting to buy something in London, Sterling is sought, rather than Euros. Places are shared spaces of action and so they come with a kind of a logical structure. Reasons can come in the sense of logical entailment or pragmatic presupposition, or more generalised warranty considerations regarding the sincerity and legitimacy of actions, among other things. I can command you to do something only if I'm warranted by being suitably superior in some regard. You can trust in my promise only if I am judged to sincerely undertake my obligation. Such warranty considerations occur within contexts like that of truth-preservation, wherein logical relations are of central importance, but operate on a less parsimonious conception of reasons than logical premises and inference rules. These reasons concern truth, truthfulness and normative rightness.

Assertions are obviously linked with truth, expressives with truthfulness (sincerity), and commands (etc.) with normative rightness (legitimacy or accountability). These three ways in which queries can be raised in conversation are themselves raised in Habermas' discussions of 'the validity basis of speech' as characterised in the late 1970s [13, p119].

---

[4] See, for instance, Heidegger, M. (1996) *Being and Time*, trans Joan Stambaugh, Albany: State University of New York Press § 32, pp. 140-41

The validity basis of speech is based upon the fundamental thought that in the very act of uttering, a speaker is claiming to be:

- giving [the hearer] something to understand
- making himself thereby understandable; and
- coming to an understanding with another person.

These are three 'world relations' taken to be implicit in speech action; fully successful speech acts must satisfy conditions of truth, sincerity and rightness (i.e. legitimacy according to some specifiable lights). [7, p75]

In the context of the robot citizen, the important question here is whether these world relations can be enjoyed by an artefactual agent. Or again, if such relations can be enjoyed, can they be recognised? The importance of these questions will be seen to hinge on some further details of the Habermasian account, and its relation to action in public, therefore to place-making over and above the mere sharing of space.

4.3.1 Further adventures in Habermas

Habermas supposes that validity claims in each of these three spheres can be raised and redeemed in communicative encounters, which amounts to raising and redeeming claims by means of argument. This being the case, validity in spheres beyond that of truth can be thought of as involving a notion of correctness appropriate to their own standards as truth is appropriate to claims of factual accuracy.

In this context, the phrase "validity claim," as a translation of the German term Geltungsanspruch, does not have the narrow logical sense (truth-preserving argument forms), but rather connotes a richer social idea—that a claim (statement) merits the addressee's acceptance because it is justified or true in some sense, which can vary according to the sphere of validity and dialogical context. [3]

By validity claims, then, is meant symbolic or explicitly made defensible propositions, sensitive to context; "A validity claim is equivalent to the assertion that the conditions for the validity of an utterance are fulfilled." [7, p38]

This is not necessarily the validity in which logicians are primarily interested, but rather must be taken to include in its scope the nuanced sense utilised above. Given we are in communication and not in some way merely noting one another's utterances, or in a therapy session or some other special type of interaction, we have to expect to be able to assume boundaries that themselves engender questions clustering around these three themes. That is to say, there are features of conversational interaction *per se* that we ought to be able to rely upon as underwriting expectations that can be shared by speaker and audience alike regarding the reasons that ought to be pertinent to their utterances. Reasons come in different flavours but can in general be requested owing to queries based in truth, sincerity and accountability.

In communication, then, what is of most interest is the role of rational compulsion as opposed to any other kind of motivation, such as fear of sanctions, for instance. The rationality associated with the simple securing of aims efficiently, instrumental (means: end) rationality, Habermas calls 'cognitive-instrumental' rationality. The concepts that would redeem validity claims in this context are simply those that, presupposing some goal, would with little fuss secure that goal. Habermas says that this much is true but goes on to stress the role of the criticisability of the knowledge claims in this area that is important but often overlooked.

The instrumental account presupposes knowledge of goals, circumstances and available means toward ends. Since with respect to each of these areas of presumed knowledge we can be mistaken, other people are apt to be able to show us that we are mistaken, perhaps by pointing out something that we've overlooked about the situation, for example.

Immediately, with this recognition, communicative rationality has expanded beyond the boundaries of mere means: end rationality. Now included in the list of presumed knowledge is knowledge of goals, circumstances, means toward ends and reasons and consequences (or ramifications). In short, the recognition of the criticisability of some presumed position opens a horizon for fallible propositional knowledge.

In acts of assertion, Habermas believes, the same knowledge is put to work as in teleological reasoning, but in a significantly different way. In action aimed at some goal, the actor can assess the rationality of their action alone and in silence. A criticisable assertion on the other hand must be rationally appraised in communication. It must be backed up with reasons or shown to be baseless with reference to another speaker's assertions in a public space of reasons.

A further extension of this simple realisation allows more candidates for rational appraisal than actions and assertions. If propositional contents can be rationally appraised on the basis of the redemption of the validity claims they raise, then other classes of expression too will be capable of rational appraisal based in the validity of the claims that they raise.

> In contexts of communicative action, we will call someone rational not only if he is able to put forward an assertion and, when criticized, to provide grounds for it by pointing to appropriate evidence, but also if he is following an established norm and is able, when criticized, to justify his action by explicating the given situation in the light of legitimate expectations. We even call someone rational if he makes known a desire or intention, expresses a feeling or a mood, shares a secret, confesses a deed etc. and is then able to reassure critics in regard to the revealed experience by drawing practical consequences from it and behaving consistently thereafter. [7, p15]

Thus, for an account of rationality larger than the mere means: end variety, we have to consider intersubjective communication in all its familiar forms. An intersubjective account of rationality has to include the possibility of validity of spheres such as those of sincerity, truth, efficacy, appropriateness, legitimacy etc. since these constitute real parts of communication. The spectrum along which validity claims can be raised and redeemed is thus much wider than merely goal-directed action and assertion.

4.3.2 Citizen-rationality and a public space of reasons

For any putative public agent, and so any citizen, this world of reason-giving and critique is a *sine qua non*. It is so owing to the requirement that place-making involves the holistic features noted from Mediterranean studies and the notion of public reasons. This is relevant to place-making as place-making requires taking an interest in the environment as a rationally structured space of

reasons, as outlined in Searle's position. The critical potential contained within the dialogical account of citizen-rationality being outlined here makes the notion of acceptability important. Acceptability, in short, must be a reasoned acceptance, not an external determination, by a citizen of a norm, value, rule or what have you.

With this idea fleshed out by drawing upon Searle and Habermas, on the rationale given by Aristotelian and Kantian thought, we now have an account of what citizens ought to be thought of as taking an interest in (institutional reality and social ontology). We also have a way of understanding how they might take such an interest (Habermas' validity-theoretic account). What remains to be explored are the conditions that would have to obtain in order to grant access to this citizen-rationality and social institutional reality. In exploring this, some of the rhetorical questions raise so far can begin to be addressed.

5. Robot access to validity spheres?

The know-how brought to bear in being able to navigate the various contexts in which citizens routinely operate is *embedded* within the world in which they arise. Social action and social performances depend on facts about the world around us, including other citizens who are themselves *enacting* the institutional and social aspects of that world. That these terms arise in the manner that they do is interesting. Given so much of place-making (status functions, institutional reality, the Background, civic nous) is concerned with the contingencies of getting on in a shared, reason-providing environment which is enacted by those who inhabit it, it is highly probable that *embodiment* is central here too. This will be explored by way of the so-called '4Es' programme. This will begin to answer the questions raised above concerning the possibility, the impetus, that a robot could have for taking an interest in public life. This will be a beginning to understanding the conditions that would need to obtain for the robot to be understood as a place-making citizen.

5.1 Embodiment

Drawing upon the '4Es' research paradigm, it is possible to gloss a few relatively recent developments in thinking about cognition. These developments suggest that cognition is:

- Extended
    - the material vehicles underpinning cognitive states and processes can extend beyond the boundaries of the cognizing organism.
- Enactive
    - It depends on aspects of the activity of the cognizing organism
- Embodied
    - cognitive properties and performances can crucially depend on facts about our embodiment
- Embedded
    - cognitive properties and performances can crucially depend on facts about our relationship to the surrounding environment[5]

While these are intended to be read as insights to cognition, they can be deployed here in the context of this discussion of the robot citizen. The following sections will make the necessary connections to demonstrate this.

Thinking about the mere space-dweller, we can easily see parallels with various artefacts. For example, we might think of a robot mapping its environment by means of measuring paths of free travel and plotting obstacles so as to come to a geometry or a topography of the immediate area.[6] Were we to anthropomorphise here we could suppose the robot to be interested only in empirical truths concerning the environment. In considering the possibility of an artefactual citizen, however, it has to be asked whether and how a robot, programme or machine could get on with place-making.

This is now the opportunity to begin addressing the rhetorical questions raised earlier, *viz*:

- From where might the impetus to get involved in governance arise for a robot?

And

- Could the robot be taken for something capable of ascribing status functions on one thing or another? To what extent could the robot *enact* institutional and social reality?

What would need to be the case for a robot to meet the criteria for being a place-maker? Much of place-making is concerned with the contingencies of taking an interest in a shared environment, it

[5] Adapted from Ward, D., Stapleton, M., https://www.academia.edu/648508/Es_are_Good_Cognition_as_Enacted_Embodied_Embedded_Affective_and_Extended (November 2011):

"...the material vehicles underpinning cognitive states and processes can extend beyond the boundaries of the cognizing organism (Clark & Chalmers, 1998; Hurley, 1998; Clark, 2008).Cognition is enactive – that is, dependent on aspects of the activity of the cognizing organism (Varela, Thompson & Rosch, 1991; Hurley, 1998; Noë, 2004; Thompson 2007). Cognition is embodied – our cognitive properties and performances can crucially depend on facts about our embodiment (Haugeland, 1998; Clark, 1997; Gallagher, 2000). Cognition is embedded – our cognitive properties and performances can crucially depend on facts about our relationship to the surrounding environment (Haugeland, 1998; Clark, 1997; Hurley, 1998,). Finally, cognition is affective (Colombetti, 2007; Ratcliffe, 2009) – that is, intimately dependent upon the value of the object of cognition to the cognizer."

[6] For a brief overview see Thrun, S., *Robotic Mapping: A Survey*, http://robots.stanford.edu/papers/thrun.mapping-tr.pdf, 2002

seems likely that *embodiment* is central here.[7] Were citizens to be each of radically differing physical forms, the emergence of an institutional reality would not be clearly of interest to any particular individual. Nor might such an emergence be possible — what might constitute general social norms for groups so diverse as to have radically different vulnerabilities and strengths?

Were citizens to be each of *radically* differing physical forms, the emergence of an institutional reality would not be clearly of interest to any particular individual. Where height, say, ranged randomly from millimetres to hundreds of meters, little sense could be made for, say, urban planning. Could a robot embodied as a dense, cubic kilometre of titanium, regardless of its faculties or apparent consciousness, possibly be understood as having interests in the environment comparable to putative fellow citizens? Or again, where a robot was embodied as a vast network of informational nods, ranging across galaxies, with an emergent consciousness, what sense could anyone make of it as a fellow burgher? It seems unlikely that such cases would permit the sort of *sensus communis* reasoning from Kant, or a comprehension of what validity claims could arise for such a being.

Another way in which embodiment reveals itself to be important in this context is in terms of the linguistic foundation to institutional reality that Searle points to and that Habermas elaborates. We can think of money as a promise, for example. Sterling notes actually state explicitly that they are promises from the bank to 'pay the bearer on demand.' Status functions in general are declared into existence and remain in existence through collective acknowledgment. The Background too can be seen as importantly linguistic, as the set of possible or counterfactual, justifications one would have given for an otherwise wordlessly performed act.

The particular way in which social beings are embodied plays a role in how and why they assign status functions the way they do, and so the institutional reality in which they act. The Background informs their mutual interactions like the terrain informs the way someone walks around. Civic nous has the impact on social action it does because it matters that another's social actions ought to be able to be anticipated and so personal actions not be perpetually frustrated.

Similarly with the case of the Background and civic nous, nowhere in particular is there a locus of this knowledge. There is a generalised pervasion of nudges, sways, insights and hints that constitute civic knowledge, that is, the knowledge of how to traverse institutional reality. From politeness on escalators, queuing for buses, paying bills, ordering beers in bars... laws, customs, habits, practices are nowhere codified once-and-for-all but rather they are more or less in any scenario to the extent that any given action is open to criticism or praise on how it matches up to this non-linear set of things.

Any artefact would seem therefore to need to be embodied in a comparable way to its social counterparts if it was to be considerable as a citizen. Any robot citizen would very likely need to be on a generally humanoid scale, with vulnerabilities similar to those of other humans.[8]

If civic, social or institutional reality is *enacted* by those whose relevant cognitive ability is *embodied* and reliant on being *embedded* amid details of the environment, then it is *extended*. The fact that this reality is extended makes it clear that it is public and up for grabs in a public way. No amount of navel-gazing reflection can arrive at a definition of what counts as this reality or its proper participants. Using the concept of recognition, we turn now to this last point.

5.2 Recognition

Could a public really detach itself from the view of the robot as servant? Could any given social group genuinely come to perceive the actions of robots as free in a robust sense? Given what has been said about embodiment just now, it could be guessed that a humanoid robot would stand a better chance than something thoroughly unlike a human in appearance. But it might also be guessed that even the most human-like robot would see diminished esteem upon the revelation of its artefactual nature. These are empirical questions, and themselves internally complex (i.e. is the possibility in question logical, practical, psychological etc.)

If the answers came in the negative, regardless of the actual capacities of the robot, none could ever be anything but a subject of oppression. Where recognition of agency is missing, there could be no chance of a full exercise of that very agency. In the republican senses of citizenship above this is the case owing to the unrecognised being unfree to take part in civic life, governance or the life of the community. In terms of Searle, the problem would be the robot not being taken as capable of enacting social reality. The suspicion of human citizens might be that the robot isn't experiencing 'the background' as the riverbed to their stream of action. Rather, something inauthentic might be suspected – behaviour in accordance with social norms read as rules. At best, the robot in these circumstances could enjoy only qualified citizenship, at worst be deemed imposter.

Bryson provides a perspective on robot identity that presses this negative line. [4] In this view, the robot is always, no matter how it is realised, an artefact directed by, and for the use of, human beings. The argument for this includes the claim that since human beings design, manufacture, own and operate robots, these robots are entirely the responsibility of human beings. This places them at the disposal of human beings, with at most the status of servant. Under no circumstances ought personhood or anything like it be attributed to the robot, on Bryson's analysis. To make such attributions would be to distribute incorrectly responsibilities and resources.

Certainly, in the area of interpersonal relationships this would be deeply problematic. Where a companion is sought, in the sense of

[7] In terms of robot rights embodiment arises too. See for instance, a discussion on 'building in' ethics to robots mentioning humanoid forms and interactivity at:

http://link.springer.com/chapter/10.1007/978-4-431-54159-2_14. Also, in

http://www.i-r-i-e.net/inhalt/006/006_Veruggio_Operto.pdf

p.3 especially, the humanoid form is mentioned.

[8] One could imagine the argument running for other types of being in a similar way, such that humanoid scale mammals or artefacts would be problematic for them. Ditto softbots. The provision of a typology isn't the focus here, but could be a very interesting undertaking.

a friend or partner, the freedom of the other is a necessary condition. Where that freedom is diminished in some way, relationships are possible but from a narrower base of, say, functional interdependencies. In the absence of robot freedom, robot companionship beyond such an interdependence is a non-starter.[9]

5.2.1 Servant machines

Bryson (*ibid*) offers a position paper and, perhaps as a result, the argument is somewhat unclear. A fourfold condition is deployed to underwrite the properly servile nature of the robot. The design, manufacture, owning and operation of robots raise different issues, especially with respect to responsibility. For example, where a robot's behaviour leads to, say, personal injury it is an open question as to whether the responsibility for this lies in the design, manufacture, ownership or operation of the robot. Whether the designer, manufacturer, owner or operator is to take the blame for the bad outcomes is a serious question with potentially very high stakes.[10]

If the position stated in Bryson doesn't exhibit a genetic fallacy, discounting robot freedom on the basis of robot origins alone, its soundness might still be questioned. The part of the argument presented here[11] that robots cannot be more than servants states that:

1.) nothing designed, manufactured, owned and operated by human beings can be anything but for our use

2.) robots are designed, manufactured, owned and operated by human beings

3.) robots cannot be anything but for our use

Whilst this is a valid argument as it stands, assumption 1 seems to be controversial. A tremendous literature and research culture exists precisely to investigate the issues that would verify or falsify the proposition. It seems too quick to rely on this as assumption when much of what is at issue is contained within the very proposition. In fact, assumption 1 seems like a *refusal to recognise* robots as having a status beyond servant.

In fact, it seems evident that no matter the success or failure of the research programme aimed at clarifying the notions of assumption 1, it is not a guarantee that human beings would accept or reject robots as more than servants. The recognition of robots as citizens, or of any *x* as *y*, would in part involve what non-robots are willing to recognise as social or political involvement.

As has been argued elsewhere (in a different context), this cuts both ways. In the same way that machines could possibly be recognised as members of a community, "...so too might an unquestionably facultative being, of silicon, carbon, or anything else, be *excluded* or unrecognised where no such well of esteem exists." [6]

The refusal to recognise as valid institutional or social action subverts the status of the putative actor *regardless of their innate nature*. Action in context, recognised as such, is central to ascribing citizenship. From Searle's account, this active, context-sensitive dynamism is clear. Building upon Searle's account and drawing upon arguments above and the 4Es programme, it is possible to make a suggestion as to what would need to be the case for a robot to be recognised as a citizen:

Where the robot is *embodied* such that it has interests in the nature of public space, it can be considered as capable of taking part in social cognition *embedded* in details of the environment. In this context, it could be possible to recognise the robot as *enacting* various institutional or social roles that could constitute or enrich this *embedded* social cognition. The interplay of these extrinsic factors, open to recognition or not, would demonstrate the *extended* nature of institutional and social reality.

6. Conclusion

This paper pursued the following strategy in exploring what would need ot be the case for the possibility of a robot citizen:

It discussed citizenship in general terms, drawing upon a notion of 'taking an interest' and substantiated this with reference to Aristotle. An absence of dependence upon power is used in Aristotle as a *sine qua non* for strict, unqualified citizenship. Kant provided an even more general means of understanding the need for other-directed reflection where agency is at stake. Drawing upon Kant's account to make a political agenda, there is the sense that reason ought to constrain power, as private and public reason are contrasted. Between these two thinkers, a view of the individual and community is advanced, with a central place for freedom and reason.

The argument then discussed *in what* an interest should be taken, by the nascent citizen (once more in abstract terms). This was the 'shape' of institutional or social reality and this contextualised in a civic setting the sort of free and other-directed reasoning seen in the first step. Searle and Habermas provided material here which provided the objects for civic reasoning, but access to these object, or to this context, for the robot remained unresolved. How the robot citizen could gain this access was discussed in terms of embodiment, and the associated notion of recognition.

For the robot to be considered a citizen there is an onus on non-robots to recognise a robot citizen – robots can't be thought of as mere objects subject to arbitrary power. This is no small undertaking, especially when it is considered that many human beings still refuse such recognition for other human beings. An essential part of gaining recognition is the embodiment of the robot citizen

It was shown that embodiment was not just a simple device to garner esteem through mutual likeness between robot and non-robot. Rather, embodiment opens doors to enactivism, embedded social cognition and it acknowledges the extended nature of institutional and social reality. It provides a way in which to understand how things can come to matter to the robot citizen as

---

[9] And so the EU programme already mentioned would be a misguided novelty cf. http://www.robotcompanions.eu/

[10] See, for instance, the case of military robots: Taddeo, M., 'Information Warfare: A Philosophical Perspective' In *Philosophy & Technology*, March 2012, Volume 25, Issue 1, pp 105-120, 28 Jul 2011

[11] This isn't the only argument presented in Bryson's paper. An extended mind position is presented, for instance, urging a la Chalmers that robots can be thought of as extensions of our own minds. Perhaps so, but the assertion is too strong in being context-insensitive: friends, relatives, enemies and strangers could all be so thought of in the right context. It doesn't determine that robots can at most be servants.

they might matter to the non-robot citizen. It is a way in which the robot can be thought of as *taking an interest*. This lays the groundwork for the possibility of place-making beyond mere space-sharing, hence of citizenship.

7. ACKNOWLEDGMENTS

Many thanks to the ETHICOMP Panel reviewers, and the membership of the Centre for Computing and Social Responsibility at De Montfort University. Their helpful comments transformed earlier drafts of this paper.

8. REFERENCES

[1] Agnew, J, 2011 'Space and Place' in Agnew, J, and Livingstone D, (eds.) *Handbook of Geographical Knowledge*. London: Sage, 2011, extract at

http://www.sscnet.ucla.edu/geog/downloads/856/416.pdf

[2] Aristotle, *Politics*, Book III, in McKeon, R., (Ed.) *The Basic Works of Aristotle*, Random House, NY, 1941

[3] Bohman, J, Rehg, W, 'Jurgen Habermas', in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), http://plato.stanford.edu/archives/spr2008/entries/habermas/

[4] Bryson, J J, 2010 'Robots Should Be Slaves' in *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issue*s, Yorick Wilks (ed.), John Benjamins

[5] Dallmayr, F, 2001, *Achieving Our World: Toward a Global and Plural Democracy*, Rowman & Littlefield Publishers

[6] Erden, Y J, & Rainey, S, 2012 'Turing and the Real Girl', *New Bioethics: A Multidisciplinary Journal of Biotechnology and the Body* 18 (2):133-144

[7] Habermas, J, 1984 *The Theory of Communicative Action*, Vol. I, Polity Press

[8] Habermas, J, 1996 'What Is Universal Pragmatics?' in *The Habermas Reader*, (Outhwaite, W, Ed), Polity Press

[9] Kant,I, 1784, 'Idea for a Universal History from a Cosmopolitan Point of View', Translation by Lewis White Beck in Beck, L, W, 1963, *Immanuel Kant, On History*, The Bobbs-Merrill Co.

[10] Kant, I, 1784, An Answer to the Question: "What is Enlightenment?", https://web.cn.edu/kwheeler/documents/What_is_Enlightenment.pdf

[11] Kant, I, 1892 *Kritik of Judgement*, Translation by Bernard, J. H, London:Macmillan & Co

[12] Searle, J, 1983 *Intentionality: An Essay in the Philosophy of Mind* New York, Cambridge University Press

[13] Searle, J, 1995 *The Construction of Social Reality*, New York, Free Press

[14] Searle, J, 2001, *Rationality in Action*, Cambridge, MIT Press