Accepted Manuscript.
Book chapter (https://doi.org/10.1007/978-3-319-97550-4_12) published in “Cognitive Architectures” (https://doi.org/10.1007/978-3-319-97550-4), Springer Nature, 04/09/2018.

Non-human intention and meaning-making: an ecological theory

Michael A.R. Biggs
University of Hertfordshire, UK
email: m.a.biggs@herts.ac.uk

Abstract

Social robots have the potential to problematize many attributes that have previously being considered in philosophical discourse to be unique to human beings. Thus if one construes the explicit programming of robots as constituting specific objectives and the overall design and structure of AI as having aims, in the sense of embedded directives, one might conclude that social robots are motivated to fulfil these objectives and therefore act intentionally towards fulfilling those goals. The purpose of this paper is to consider the impact of this description of social robotics on traditional notions of intention and meaning-making, and in particular to link meaning-making to a social ecology that is being impacted by the presence of social robots. To the extent that intelligent non-human agents are occupying our world alongside us, this paper suggests there is no benefit in differentiating them from human agents because they are actively changing the context that we share with them and therefore influencing our meaning-making like any other agent. This is not suggested as some kind of Turing Test in which we can no longer differentiate between humans and robots, but rather to observe that the argument in which human agency is defined in terms of free will, motivation, and intention can equally be used as a description of the agency of social robots. Furthermore, all this occurs within a shared context in which the actions of the human impinge upon the non-human and vice versa, thereby problematising Anscombe's classic account of intention.

Introduction

One way to describe human beings is as meaning-making agents. What we do is to interact with the world and our interactions are mediated by the meanings that we make. It is no longer fashionable to think that those meanings are inherent in the world around us, as was implied in classical hermeneutics, but rather that we project onto the world – in its infinite complexity – our interests and motivations which organise themselves as meanings that we commonly say we "find" in the world. Thus one can describe human beings are active agents, which is indeed the premise of actor network theory. When Latour wrote about scientists in the laboratory (Latour and Woolgar 1986) he observed their activities as a kind of interested behaviour in which they found meaning in scientific activities owing to their identity as scientists.

Accepted Manuscript.
Book chapter (https://doi.org/10.1007/978-3-319-97550-4_12) published in “Cognitive Architectures” (https://doi.org/10.1007/978-3-319-97550-4), Springer Nature, 04/09/2018.

This contemporary view of human beings as active agents and meaning-makers assumes that the external world is largely passive in response to our activity. However, intelligent tools increasingly accompany the activity of human beings, and these tools often mediate our interactions with the world. Thus when we use Google to search for information, in addition to our motivation to direct the search engine by an informed choice of keywords, Google itself is operating according to algorithms and protocols that guide searches in one direction or another. Of course these robots are there to do our bidding, but as they become increasingly intelligent our world becomes populated by artificial versions of ourselves. If we have motivations and interests that are determined by our overall aims and objectives, can we not say that these robots also have motivations since they too have aims and objectives inherent in their programming? When we proliferate the existence of these robots more widely in society, where forms of artificial intelligence [AI] mediate so many of our interactions, are not our naïve assumptions that we alone are the active agents no longer viable?

This is a profound change Anscombe's (1965) notions of intention, motivation and responsibility. It also tends to equalise the relationship between the human user and the robot so that the human must take account of the capacities and interests of the robot when engaging with it, and as such our sense of meaning in the world must now include other active, inorganic agents with whom we must negotiate our own meaning-making activity. Conversely, if we can adopt a somewhat anthropomorphising thought experiment, could one not say that the robot was equally engaged in a meaning-making activity in which negotiating with humans was necessary?

Meaning making has hitherto been discussed as an essentially human activity. It is an interpretative act that we employ as part of optimising our agency in the world. When we act, we generally act with an aim in mind, and so our actions must be designed to have a certain effect and to overcome potential barriers or resistance. Thus we need to understand the context in which our actions will take place and the factors that may impinge on them. This interpretation of the ecology of meaning making includes understanding what's going on in the current situation so as to intervene effectively to achieve the new and desired situation. All this is normally described in terms of intention to achieve something, based on an interpretation of what would be necessary and the exercise of will that brings about transformation. But intention alone is not enough to bring about effective change because it must be accompanied by an effective understanding of what must be done to meet the intended outcome. In an AI environment this is the difference between intelligent and non-intelligent robots. The non intelligent robot may have an “intention”, that is to say “programming” to do something but if the environmental factors are not as expected, for example, an object is not where it is supposed to be, the non intelligent robot cannot achieve its intention, i.e. the programmed goal. In such a simple example, the meaning of the situation is that a different set of actions to that originally programmed is necessary to achieve the intention. In more

Accepted Manuscript.
Book chapter (https://doi.org/10.1007/978-3-319-97550-4_12) published in “Cognitive Architectures” (https://doi.org/10.1007/978-3-319-97550-4), Springer Nature, 04/09/2018.

complex social environments, meaning may consist in identifying other dynamic agents in the environment, or hypothesising their intentions and hence predicting their behaviour. These new possibilities problematize our existing notions of intention and meaning making because hitherto they have been seen as essentially human traits that to some extent differentiate humans from machines. This paper considers whether such terms should continue to be reserved for humans or whether recent developments in AI should cause us to reassess our understanding of intention and meaning making as something environmentally situated or ecological rather than individually situated and subjective.

An ecological theory

The traditional approach in philosophy has been to differentiate humans from non-humans, including intelligent machines and robots. Although both humans and robots have agency and can interact with the material world and change it, we have aggregated to the concept of "human" some superior powers such as free will, interpretation and intention. Under this approach, robots do not have the capacity to exercise these essentially human qualities and therefore have no responsibility for what they do. Free will, interpretation and intention have each been the subject of extensive analysis in philosophy, and the notion of intention has been examined in detail by Anscombe (1965). In her classic paper she identified three different kinds of intention: (1) intentions to act, (2) intention in acting, and (3) intentional action. Intention is closely related to the idea of choice and of will (volition) in which the human agent brings something about and can be said to be responsible for it, both in terms of causality and moral responsibility. Although we recognise that robots can have agency and be causally responsible for change, as in the case in which a factory robot builds a car and is certainly the agent of change that brings about its construction, we do not normally speak in terms of the moral responsibility of robot. The responsibility for the robot's actions, if brought to a court of law, would probably be found to lie with the programmer because the robot “mindlessly” carries out the instructions that have been given to it. But why should we make these distinctions? Although we might desire that human beings be differentiated from other animals and from inorganic actors, in the world of artificial intelligence [AI], can, should, ought such differentiations be sustained? Indeed what would be the point?

When we intend to do something, an aspect of that intention is that a future plan may or may not be realised. This is discussed in Anscombe’s first category of intention. On the one hand it may not be realised because we change our mind and we do not act as we originally intended. On the other hand we might act unsuccessfully and not bring about what we intended. In either case, the future prediction embodied in the intention did not happen but we nonetheless say that there was a motivation so to act or to bring something about. One of the things that we expect about robots is that they will successfully bring about what they

are programmed to do, over and over again. Furthermore, robots are not usually regarded as having the capability of changing their minds in relation to this behaviour, if we regard “what is in their mind” i.e. what we might informally call their “intention”, as being embodied in their initial programming. Of course, AI allows adaptation but this is probably not the adaptation of an overall aim even if the adaptive system may have the capacity to "change its mind" about how to achieve that aim. Thus the changing of intermediate objectives as an apparent expression of the “intention” to fulfil an overall course of action, turns out to be something that could be meaningfully referred to in relation to inorganic agents such as robots, as well as organic agents such as humans.

We sometimes have the intention of bringing about A, but inadvertently, we bring about B. Although B was brought about, we cannot say in good faith that we had the intention of bringing about B, although we are sometimes disingenuous, and in order to save face, we say, “I meant to do that”. “Meaning to do something” is an utterance, not a speech act. That is, just saying “I meant to do that” does not make it so. When I intend to bring about A, I may say aloud in advance that I predict this will happen, or I may make purposive actions that under normal circumstances would bring about A, or it may be assumed by myself and perhaps others that I am attempting to bring about A on the grounds of my past history or the perception of my interests by others. But the mere subsequent utterance of the statement "I meant to do B" does not mean that, after all, I really intended to do B rather than A. Such an utterance would be regarded as post-rationalisation in psychoanalysis, and face-saving in negotiations. As yet we have not deemed it necessary to programme face-saving into robotic behaviour. Thus adaptive behaviour, in humans and in robots, should normally not involve a change in the overall goal, only the means of achieving it, i.e. of changing the intermediate goals when necessary.

So the question remains, is it useful to say that robots have intention even though they do not say "I meant to do that" and furthermore, what would be the consequences of this change of attitude in the case of less evident robotic agents such as social robots, which operate more discreetly at the margins of our environmental awareness, if they were said to have intention? In other words, is intentionality something that I attribute to an agent when I see indicators of "acting to bring about", or should intention imply the possibility of failure that is normally missing from programmed robotic behaviours? Indeed, is intention so inherently human -- because to err is human -- that it is meaningless to speak of a robot's intentions when they are always satisfied? Conversely, is it essential that we keep open the possibility of an intentional act in order to attribute responsibility for action, and to whom and under what circumstances should that responsibility be attributed to the human programmer or the robotic actor? This is the problem discussed by Anscombe in her third category; intentional actions.

One of the “traditional” assumptions that would be problematized is the so-called Turing Test (originally framed as "can machines think?", or "exhibit

Accepted Manuscript.
Book chapter (https://doi.org/10.1007/978-3-319-97550-4_12) published in “Cognitive Architectures” (https://doi.org/10.1007/978-3-319-97550-4), Springer Nature, 04/09/2018.

intelligent behaviour", Turing 1950: 433 & 459) in which it is proposed that differentiating between robots and humans would be irrelevant if we could not differentiate one from the other through questioning. This procedure assumes that we have an explicit encounter with another being whom we suspect maybe a robot in a situation where we might normally expect to meet a human, or vice versa. The scenario posited in this paper is slightly different. The scenario is that we are frequently confronted with social robots that are intelligent agents seamlessly integrated into our social environment. If we posit a seamless interaction then we cannot know whether this agent is human or not, and the issues of intention and responsibility are indeterminable. Such a question is only likely to arise when there are questions of responsibility regarding the actions of a robot or the consequences of its action, for example by misleading a human into taking certain actions that otherwise it would not have taken. Of course, as has already been described, in such circumstances we might hold to account the programmer who wrote the programme that caused the robot to make the decisions that it made, leading to the undesired consequences for which the question of responsibility is an issue. But what if the social robot is integrated to a much greater extent into our social interaction so that it passes the Turing Test? And what if the robot has such a complex AI that we cannot reasonably hold the programmer accountable for the decisions that the robot has made based upon the fundamental principles embodied in its initial programming? Do we need a model for this kind of autonomously learnt behaviour?

In human society we already have a model disclaimer for responsibility in that we do not hold minors responsible for their decisions and actions. Parents or guardians, that is to say, the societal programmers, are normally held responsible until the minor reaches a certain age. It is interesting to note that the test for legal responsibility is not a performance criterion, as is the case with the Turing Test, but merely an age criterion. If we applied such reasoning to social robots we might conclude that when they had been acting autonomously in the social environment for a certain period of time, during which they evolved their AI to address most of the commonly encountered problems for which they were programmed, we might infer that they could be held accountable for their intentional actions. But this would bring us back to the earlier observation that we do not at present have the legal framework or practice of holding machines to be responsible for the actions they perform or their consequences.

However, the responsibility of the mindless factory robot is not the principal focus. Instead this paper is interested in the extent to which the concept of an, in practice, transparent agent, by which I mean an agent that cannot be differentiated from the human -- not because it passes the Turing Test but because the context in which we engage with it does not invite that kind of differentiation -- has impact on how we have previously described intentionality. Anscombe's category 1 remains apparently unchanged because there is no need to infer that artificial agents have a predictive capability. However, the ability of such agents to make change, and by their adaptive behaviour be said to assume "responsibility" for actions that were not or could not have been anticipated in

Accepted Manuscript.
Book chapter (https://doi.org/10.1007/978-3-319-97550-4_12) published in “Cognitive Architectures” (https://doi.org/10.1007/978-3-319-97550-4), Springer Nature, 04/09/2018.

the original programming, does seem to imply that we can meaningfully speak of such artificial agents as having intention. This has hitherto been assumed to be a uniquely human attribute.

Of course, historically, many of the uniquely human attributes that anthropologists have identified such as the ability to use tools, and that sociologists have identified such as the ability to use language, etc., have been proposed in order to meet the desire to differentiate human from animal. One might regard the Turing Test as the last vestige of this historical attempt to desperately maintain such a teleological differentiation. But if we abandon our attempt to be different, in addition to the practical issue of whether we can still make such a differentiation or indeed whether it is necessary or productive, what would be the consequences of believing that artificial agents can have intention?

When agents act intentionally, in the sense of Anscombe's category 1, we attribute some kind of motivation or plan to them. We say they want to bring about outcome X. In this scenario I am not merely thinking of machines with direct programming in which we can say they “mindlessly” act to bring about outcome X and so they themselves do not meaningfully have an intention. In the present scenario I am assuming AI of sufficient order, coupled with social embeddedness that renders the agents invisible, that we are unaware of the human/machine distinction and are only aware of the agency, the purposiveness, and the responsibility. Having granted a category of inorganic intentionality, which is perhaps additional to Anscombe's original three, what does this tell us about the “inner life” of these inorganic agents, and about this extended concept of intention? Do they feel satisfaction when their intentions are fulfilled? Do they feel frustration when they are not? Do they have an overall perception of the environment in which they are operating within which they frame their decisions according to their programming and subsequent experience?

To all these questions it would be most interesting to answer yes. Yes, they do have responsibility for their actions; yes, they do feel frustration when their intentions are not satisfied; and yes, they do have an overall perception of the environment in which they are operating.

This is not merely a science fiction discussion in which we ask whether androids dream of electric sheep: it is a philosophical discussion about the consequences of integrating social robots that act intelligently into the human environment, and to ask how to attribute intentionality and responsibility when we interact with them. There is a reason why we should be interested in this problem. In contemporary philosophy the focus has shifted from ontological and teleological issues in which the question or the questioner is to some extent independent of the social environment, to questions that recognise social ecology and relational judgements, and worldviews that require meaning-making. Relational judgements imply that if we are sharing our world with other agents, whether

Accepted Manuscript.
Book chapter (https://doi.org/10.1007/978-3-319-97550-4_12) published in “Cognitive Architectures” (https://doi.org/10.1007/978-3-319-97550-4), Springer Nature, 04/09/2018.

they are organic animal agents or inorganic robotic agents, the network of relationships will present certain possibilities. Therefore it is relevant to know who and what is in our environment and the way in which the other, owing to being dynamic, is causing change to our environment and therefore the decisions that we make. At a macro level this means that our worldview is impacted by our perception of what is material, of what can change, and who are the agents of change that are independent from us. Furthermore, we have to take account of the apparent worldview of those change-making agents in order to predict their behaviour that may impinge upon us and our ability to successfully implement our own intentions.

So we have perhaps arrived at the possibility that artificial agents, owing to their capacity to act dynamically and to impinge upon our worldview, can clearly be said to themselves have intention of category 1. This argument also suggests the possibility that these intelligent agents are making meaning for themselves so as to fit their actions to the environment. In the past, meaning-making would also have been an ability reserved for the human. But if we abandon our differentiated status we can now see that these intentional acts that are based upon the experience gained by the social robot that is embedded in the same environment as ourselves, are inevitably based upon the same decision-making structures as ourselves. Indeed, causally, the decision-making actions and strategies of the inorganic agent were brought about through its programming by a human agent. The inorganic agent, in this case the seamlessly embedded social robot, however modest it may be in its capacities, is brought up as a minor with strict instructions in its programming that determine its behaviour. During its formative years of operation it develops, through the use of its intelligence, experiences and additional frameworks that enable it to make decisions that were not framed or anticipated in its original programming. This is what we want the intelligent robot to do when we design it -- so that it is unnecessary for us to anticipate every possible scenario in which it must take action and to determine the action it must take. An effective social robot must be judged responsible for the decisions that it takes because it has taken them based on frameworks of judgement that were not placed there by the programmer. The programmer is innocent, or at best merely an accomplice! To make such judgements, the inorganic agent must "understand" its environment and have a worldview. Of course, such a worldview need only stretch as far as the scope of agency envisaged for that robot. However, having postulated the possibility that one can describe the agent in this way, one has to conclude that meaning-making is being undertaken when the robot evaluates a scenario and identifies within it the possibility for action. Such a possibility for action is implied in the concept of intention because we cannot meaningfully speak of intention in the circumstance in which the intended outcome is unlikely to come about or when such an outcome would be impossible. If I intend B in the circumstance in which B could not possibly happen then my intention will be described as folly. Misguided intention, i.e. folly, is noticeably absent from Anscombe's three categories.

Accepted Manuscript.
Book chapter (https://doi.org/10.1007/978-3-319-97550-4_12) published in “Cognitive Architectures” (https://doi.org/10.1007/978-3-319-97550-4), Springer Nature, 04/09/2018.

As a result of the foregoing argument we have the possibility that now, or in the near future, we will share our environment with social robots albeit with modest remits, and that these agents will not differentiate themselves from other active agents in that environment. When we, as human agents, interact with this mixed ecology we will form a view of the active and inactive elements within it in order to frame our intentions and our actions. It is important for us, if we are not to be frustrated by the lack of fulfilment of our intentions, that we perceive the ecology in which we operate as dynamic. Relational argumentation is one contemporary outcome of the recognition of this need. It can be contrasted with the absolute argumentation of Newtonian mechanics in which the external world behaves passively, to the extent that it is not actively making autonomous decisions. This is no longer the case. All sorts of agents, some of which are inorganic, populate our world and some of those are making AI-led judgements about what to do in response to us at the same time as we are making intelligent judgements in response to them. As a result our worldview, that is the view of the range of possibilities that the world presents to us to either facilitate or frustrate our intentions, is modified by the presence of these robots. In his test, Turing argued against those who needed to, or saw the possibility of, differentiating robots from humans; but that possibility no longer exists, not only as a consequence of the increased adaptive intelligence of the robots but also their embedded-ness in our social world. Our ecology now includes new autonomous agents of change.

So what are the consequences of this philosophical description of robotic intention and meaning-making? We have seen that one can usefully refer to both intentionality and meaning-making in the case of robots, and that there may therefore be no need to differentiate human agents from embedded social robots because they share the same societal and legal responsibility for their actions. The Turing Test becomes irrelevant in such cases because there would be no benefit from making the distinction. The traditional distinction allows humans the exclusive right to free will and responsibility, but it seems that such a distinction is no longer beneficial or sustainable. This paper has suggested that it is an inevitable consequence of the increasing adaptive complexity of social robots and their embedded-ness in the environment in which they become part of our social ecology, that we will have to begin to deploy concepts that have previously been reserved to humans. The concept of intention is one such concept. It is meaningful to speak of the intention, whether fulfilled or not, of the robot. The robot's actions may have intended or unintended consequences. The robot, if it is to successfully negotiate dynamic obstacles to fulfilling those intentions, must anticipate -- that is to say, predict -- what will happen if it takes certain courses of action. For these operations to be successful the robot must have a worldview and must make decisions in accordance with it. Meaning-making it is perhaps the most advanced of the concepts that has been speculated here. To what extent is meaning-making really a part of the robot's behaviour?

Comparing once again to the human model, the idea of meaning-making is deployed in order to account for the way in which humans construe the world so as to anticipate how it will operate and how they can operate within it. Meaning-

making embodies the idea that there are dynamic agents in the world pursuing their own objectives, and we appear to these agents as they do to us. Their behaviour, when different from our own, can be explained by them having a different view of the world, their place in it, and them having different motives and therefore different objectives. These objectives are pursued through intentional action whether by human or nonhuman agents. Their meaning-making is not a quest for the meaning of life, but the meaning of the presence of these other agents who are not working harmoniously with their own interests. Meaning-making results in an explanatory framework that accounts for diverse interests. Now that we have created social robots to work with us, they also work amongst us, and owing to their different function have different, albeit normally harmonious, intentions to ourselves. Thus meaning-making does not emerge as an exclusively human attribute because it is linked much more to the ecological interpretation of the context in which the agent is embedded than to manifestations of some inherent subjective capability of the agent itself. Intention and meaning-making are environmental by-products of agency. This conclusion has consequences in a number of areas and for the interpretation of the previous literature.

One consequence is that Anscombe's three categories should be explored in relation to non-human as well as human agents. Intention as a concept applies in any situation in which purposive action is taken to meet an objective. This can be said to occur when there is any kind of programmed objective. One motivation for Anscombe's discussion seems to be an implied interest in responsibility, as is evidenced in her example of the man who poisons the occupants of a house (1965: §23). But equally it can be posed in relation to a robot that brings about outcomes owing to adaptive behaviour for which we cannot hold the programmer fully responsible.

With regard to meaning-making the traditional concept becomes more stretched, but it is unnecessary to hypothesise a ghost in the machine in order to find the concept of inorganic meaning-making a useful one. If we intend when we design a social robot that it should seamlessly integrate into the social context so that it can effectively serve its human masters, then we must equip it to be adaptive owing to the dynamic nature of the human context and the other human agents amongst whom in must operate. Its efficacy will be enhanced if it can not only accommodate such situations as it finds but also anticipate possible scenarios. This requires an adaptive map of possibilities that constitutes, this paper claims, a worldview. Meaning-making does not require a metaphysical conscience which gives meaning, in the sense of ultimate purpose, to the world. All that is required is the ability to project forward and anticipate so as to improve decision-making. Meaning, understood in this way as a practical activity, is making inferences from indicators. Thus when X means Y we can substitute, when X indicates Y or, as a result of X we infer that Y will come about. Put in this way it is reminiscent Hume's notion of cause and effect as merely the “constant conjunction” of X with Y (Hume 2011 [1748]). Hume's view changes the locus of causality from something that is extrinsic to perception, to something that is intrinsic, and

makes it a psychological theory. In other words when we think that X causes Y we think that something is happening in the external world. When we think that X is constantly conjoined to Y we think that something intrinsic to ourselves is happening: this is an idea rather than a fact, something that is going on inside us rather than something that is going on in the external world.

So it is with inorganic meaning-making. The concept of inorganic meaning-making is a consequence of reframing a concept that was once intrinsic so that it becomes extrinsic. When we describe meaning-making by the inorganic agent we are not giving the agent human attributes, we are simply applying the extrinsic argument. The inorganic agent can be said to make meaning when it makes the ecological connection between X and Y and adapts its behaviour accordingly. At one level this is simply predictive. At another level it confirms that the agent has a model of what will happen and is acting according to that model. Such a model consists of both objects and events, and possibilities for which there are indicators. It is the presence of adaptive behaviour in the presence of indicators that underlies the argument that meaning-making is present. The perception that X means Y in situation Z corresponds to the utterance "X means Y" and successful social robots will be deploying this concept just as frequently as do humans. Thus the context dependency of all operatives is at the root of meaning-making by any agent.

References

1. Anscombe, G. E. M. (1965). *Intention*. London, Harvard University Press
2. Hume, D. (2011 [1748]). *An Enquiry Concerning Human Understanding.* Urbana, Illinois: Project Gutenberg. Retrieved 29 July 2018, from www.gutenberg.org/ebooks/9662.
3. Latour, B. and S. Woolgar (1986). *Laboratory Life: the construction of scientific facts*. Princeton, NJ, Princeton University Press
4. Turing, A. M. (1950). "Computing Machinery and Intelligence." Mind LIX(236): 433-460