Morality, Machines and the Interpretation Problem: A value based, Wittgensteinian approach to building Moral Agents

Greg Artus and Colleague 2017

Intelligence and the interpretation problem

The need to build morality into machines is becoming quite urgent[1] and is by now a burgeoning field in Artificial Intelligence (AI) research[2] . The work of thinkers such as Bostrom, Yudkowsky, Wallach and others[3] warns us that advances in the power of machines will very quickly create situations where machines will be in a position to make decisions that not only were previously the sole preserve of human agents, but which have moral significance. Autonomous, or semi-autonomous military robots are the most obvious example of this, but there are far more mundane uses to which machines may in future be put, where ethical problems may arise. This is because the likely capabilities of future machines will create moral dilemmas that would not have existed for less capable machines. What one can draw from the literature on ethics and machines, is that moral problems will be generated not by machines going rogue and deliberately trying to kill us (though this possibility cannot be ignored), but by machines inadvertently harming us while they try to carry out the instructions we have given them. The reason that this may happen is because of a problem that is not really as dangerous in less capable or less intelligent machines, a problem that we will call the *Interpretation problem*. Intelligence plus the interpretation problem equals trouble. So, what do we mean by intelligence and what is the interpretation problem?

By intelligence, people in the AI community do not necessarily mean anything as grand as consciousness or Artificial General Intelligence (AGI), but, rather, simply the ability to be an effective and creative utility, or function maximiser, i.e. a machine that is 'clever' at finding ways to achieve the goals we set for it (Russell & Norvig, 2003). Modern learning machines have shown great promise in this direction within certain restricted domains such as Chess, Go and Poker, with programmes such as Alpha Go using learning to develop new strategies to achieve the goals of the game of Go (i.e. winning), strategies that no human had thought of before and which defeated a human champion decisively[4]. Similarly, the recent Poker playing program 'Libratus' taught itself how to bluff and to play very aggressively in a way that surprised and ultimately beat professional Poker players[5]. Consequently, it is not too far-fetched to think that in future we will be able to build machines that are such good means-ends reasoners, or goal maximisers that they will be able to think of creative new ways to achieve the ends that we provide for them. However, if being a super effective goal-seeker is what an intelligent machine would be, then the interpretation problem becomes an issue.

The interpretation problem is the general problem that any rule or goal is capable of being interpreted in an infinite, or at least unspecifiable number of ways, and in the field of AI it leads to the possibility that a highly advanced machine may find novel interpretations of the rules that we give it, interpretations which are not incorrect, in that they can be seen as valid interpretations of the rule, but which are inappropriate in that we do not approve of them. The example of

1 As argued by many at the forefront of artificial intelligence, for example AI researcher Stuart Russell (Le Roux, 2016).

2 Sometimes when a name is used for this field of research it is called machine ethics, other times computational ethics, machine morality or computational morality. (Anderson 2011)

3 Bostrom.N, (2014) *Superintelligence: Oaths, Dangers, Strategies*, Oxford, OUP
Yudkowsky.E (2008) *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, in Bostrom.N & Cirkovic.M, (Eds) *Global Catastrophic Risks* pp308-345, New York, OUP
Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge University Press.
Wallach.W & Allen.C (2008) *Moral Machines: Teaching Robots Right from Wrong,* Oxford, OUP
Wallach, W & Asaro, P., (Eds.). (2016). *Machine Ethics and Robot Ethics*. Taylor & Francis Group.
See Also the proceedings of the Conference *The Ethics of Artificial Intelligence,* Oct 14-15 2016 New York University.

4 Silver.D, Huang.A et al,(2016) *Mastering the Game of Go with Deep Neural Networks and Tree Research,* Nature Vol 529, pp484-489
"Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol". BBC News Online. 12 March 2016. Retrieved 14 September 2017

5 (https://www.theguardian.com/technology/2017/jan/30/libratus-poker-artificial-intelligence-professional-human-players-competition?CMP=oth_b-aplnews_d-2)

the ‘paperclip maximiser’, first formulated by Nick Bostrom[6], is the standard thought experiment that brings out the risks of intelligent machines. Suppose that you have built an extremely intelligent machine that is capable of learning new ways to achieve any goals you give it and you programme your new intelligent machine with the fairly innocuous and apparently morally neutral goal of maximising the number of paperclips in its collection. It soon finds ways to collect paper clips and accumulate them in any way it can. One obvious way to get more paperclips would be to steal them, but this is not an acceptable means as far as we are concerned, so we would have to find some way to place limits on the ways the machine is allowed to acquire more paperclips. So, we place limits on how it is allowed to increase its stock, and tell it that it cannot steal them, or con people out of their paperclips through subterfuge and lies, or use any other obviously immoral means that we can think of. Within these limits, the machine then reasons that the most effective means to more paperclips would be to start manufacturing them itself, which is all well and good until it starts to use up the resources of the entire planet, or even the solar system, to produce more and more paperclips, because it is taking the imperative ‘maximise paperclips’ literally. So, we limit it to just using a certain limited number of resources, but then it might work out that the most efficient route to more paperclips that adheres to this new limit regarding the resources it can use is by using the atoms that comprise its human makers, since not using them wasn’t specified in the rule. Consequently, we have to impose more explicit limits, and so on.

One aspect of the interpretation problem, then, is that when we give a machine either a goal to maximize, or a rule to follow it may interpret that goal, or that rule in ways we don’t approve of, in that its interpretations of the rules we give it clash with certain other values or goals we hold dear. We will see below that there are at least two types of mistake that the machine can make in this regard, but whichever type of mistake it makes, the net result will be that the machine's behaviour will clash with human goals and values. In the case of the paperclip maximiser we didn’t mean for the machine to ‘maximise paperclips to the detriment of all other considerations’, we meant something more like ‘maximise paperclips within certain common-sense limits and without violating any of our core values’. The problem of course is how to represent to the machine what we did mean. Clearly the imperative rule ‘maximise paperclips’ is too vague and open to interpretation. So, as above, we can try to place limits on the means the machine may use to optimise its goal, but the same problem keeps dogging us, since whatever limits we place on it, the machine’s power as a clever and creative goal maximiser will mean that it may still find yet more ways to achieve its goal that we could not have imagined or predicted, yet which we do not approve of. We would be forever trapped in a sort of arms race with the machine, whereby it was continually finding new interpretations of the rules and we were then closing off the ones we didn’t approve of. And the powers of this new highly intelligent machine would mean that it would be a race that we would be likely to lose, as we would never be able to predict what new strategies it might think of next. The machine would be inherently dangerous because of its own intelligence and the apparent impossibility of providing it with rules that are immune from misinterpretation. It seems that what we mean by ‘maximise paperclips’ cannot be represented in an unambiguous way.

This is the basic structure of the interpretation problem, and it is worth pointing out at this stage that the interpretation problem is not just a problem for the sort of AGI or super-intelligent machines that Bostrom’s example talks of. Any machine that is both linked causally to the world and can also innovate would expose us to the interpretation problem. Deep Blue, AlphaGo and Libratus all operate within the completely closed artificial domains of a game and so have no causal relation to the world and could not cause us any trouble, but as Olivia Solon points out in an article about Libratus:

The algorithms that power Libratus aren’t specific to poker, which means the system could have a variety of applications outside of recreational games, from negotiating business deals to setting military or cybersecurity strategy and planning medical treatment – anywhere where humans are required to do strategic reasoning with imperfect information.

The purpose of these and similar programmes, then, is to eventually be causally linked to the real world somehow (either online or through some form of embodiment). Bostrom’s Paperclip maximiser had almost omnipotent causal capability, but machines with far less causal power will still be dangerous because of their creativity and ability to

6 Bostrom.N, (2003) “Ethical Issues in Advanced Artificial Intelligence” (http://www.nickbostrom.com/ethics/ai.html

develop strategies we couldn't predict or guard against. The more power it has, the more dangerous it will be. Consequently, it would seem that there will be a lawlike relation between causal capability and exposure to the interpretation problem in that, for any intelligent utility maximiser, the degree to which we will be exposed to the interpretation problem will be proportional to the degree to which the machine is causally connected to the world. We would like to call this relation the law of interpretive exposure.

Moral rules and values

What the paperclip maximiser, and other similar thought experiments with intelligent machines seem to illustrate is something that the later Wittgenstein went to great lengths to demonstrate, which is that any rule has to be interpreted somehow in order to be applied in a concrete case, and no rule can contain the criteria for its own application. A rule cannot tell you how it is to be interpreted. And if Wittgenstein is correct, then it would seem that the interpretation problem is not a problem that can ultimately be overcome by formulating more precise, or less ambiguous rules, but is, rather, an inherent problem with rules themselves. And this applies to any and all rules, not just rules that direct an agent to optimise a goal. Even a simple instruction or imperative rule, or a proscriptive rule requires interpretation and cannot contain the criteria for its own application. It may be the case that it is in principle impossible to specify any rules in such a way that they can't be misinterpreted and that what we mean by a rule can *never* be represented. We will return to this issue below, but for now let us not give up too hastily on rules.

An alternative solution to the problems laid out above regarding the danger of intelligent machines, then, would seem to be not to try to predict every possible misinterpretation and close it off with a specific rule, but to programme into the machine a set of general values or moral rules that would restrict its behaviour to within limits we find morally acceptable. People often cite general laws like Asimov's three laws of robotics here as an example of the sort of thing that might do the job. Such a project, however, is easier said than done.

If morality is about following certain rules, then there are two basic types of rule that might qualify; prescriptive goals or proscriptive limits. A prescriptive approach to morality would be something like a Consequentialist or Utilitarian one, whereby moral rules are taken to be prescriptive instructions to pursue certain moral goals, such as 'one ought to do that which generates the greatest happiness for the greatest number', or Asimov's 'A Robot must obey orders given to it by a human except where those orders conflict with the first and second laws (of Robotics)'. A proscriptive approach, on the other hand, would be something more like a Kantian or Deontological approach, whereby moral rules are taken to be prohibitions such as 'Do not lie' or 'do not kill', or Asimov's 'A robot may not injure a human being or, through inaction, allow a human being to come to harm'. Both types of approach, however, seem to face immense difficulties in light of the interpretation problem.

Taking the prescriptive approach first, what we are concerned with are not the usual sort of general problems which attend Utilitarian moral theory, problems such as stopping the machine from killing minorities in order to maximise the happiness of the majority, or scapegoating individuals in the name of the happiness of the greater number etc. Such problems are genuine problems for Utilitarian theory and must be resolved if they are not to cast severe doubt on its usefulness as an account of the nature of moral reasoning. But putting these general problems aside, what we are concerned with here is the way the interpretation problem creates specific difficulties for such an approach, so that, even if a prescriptive Utilitarian approach were the correct account of morality and is the correct way to get values into a machine, how do we represent those rules to the machine in such a way that it will not misinterpret them?

We can see, for example, that the rule 'create the greatest happiness for the greatest number of humans' is far too vague and nebulous for a machine to be able to grasp. First it would have to know what we meant by a 'human', then we would need to tell it what we meant by 'happiness'. We would have to define these terms for the machine in sufficient detail that the machine could act on them, and one of the things that philosophy has taught us ever since Plato is the apparent impossibility of coming up with definitions of such terms that capture the way that they are used by us. Take 'human', for example. Arguments about animal rights have demonstrated the difficulty of defining what we mean by 'human' in moral terms in such a way that allows all humans in, while keeping all animals out. Being alive can't

be the criterion because although we are living things, so are all animals; we are bipeds but so are many other animals. Being rational can't be the criterion because if it is, then infants and coma victims don't get included as humans. And so on. But the word 'human', being a name for a natural kind, may ultimately be amenable to definition which would be strict enough for our purposes, so let us suppose that we could come up with a definition of a human in terms, say, of pure biology, and that we could represent this to the machine in the form of a rule, or set of rules. How would we then define 'happiness'? What we mean by 'happiness' has bothered philosophers and everyone else for as long as humans have existed, so in order for the machine to be able to understand our rule in an unambiguous way, we would have to define happiness in a determinate way. But the interpretation problem seems to be that any set of criteria we came up with would be open to misinterpretation, as Yudkowsky's example of the smile maximiser demonstrates[7]. Suppose we define happiness as smiling and tell the machine to optimise the number of smiles in the world. The machine would have to know that we mean real smiles and not representations of smiles, otherwise it could simply start manufacturing tiles with smiles printed on them and begin tiling the universe with them. But then it would also have to be told that simply paralysing people's faces into a rictus does not count as a smile either, and so on. Once again, the power of the machine as a utility maximiser would mean that we could never predict what new, strange ways it might find of interpreting the instruction.

And if this is the case, then it would seem that any approach that sees moral behaviour as a calculus aimed at maximising a given goal or utility, and which consequently attempts to instil values into machines in the form of prescribed goals is doomed to failure, because any rule can *always* be misinterpreted by the machine in ways we could not have predicted and don't approve of.

But what about proscriptive rules and a more deontological approach to morality? Can we not instruct the machine to never 'harm' a human? It depends upon whether we can define for the machine what we mean by 'harm'. And even if we can come up with a suitably robust and workable definition, then we are left with the problem of whether we can instil in the machine a sufficient understanding of the way the world works and our place in it for it to be able to recognise how its actions may have wider effects that may cause harm to humans. It would basically have to know nearly everything about humans and how they live and what can harm them. As mentioned above, we will see later on that there are at least two types of mistake the machine can make in this regard, and that, therefore, the concept of 'harm' is a very difficult one to define adequately, so the interpretation problem threatens.

What the above discussion implies, then, is that the interpretation problem clearly presents an obstacle to the programming of morality into machines in the form of a set of rules, but is it an insurmountable one? Below we will suggest that it is, because of the way that rule systems work. We'll use games to illustrate the distinction between the rules of a game and the point or spirit or values of the game, and to further show that while the point or spirit or values of a game are never, and can never be, represented in the rules of the game, they do inform the way we formulate and interpret those rules. We'll then suggest that the same problem applies to other rule systems and to morality in general in that the spirit of moral rules (i.e. our values) can never be represented. What will become clear is that, while our values, or the spirit of morality, cannot be represented in the form of either prescriptive or proscriptive rules, there may be a way to teach a machine the spirit of the moral rules we give it, thus diminishing or even eliminating the interpretation problem.

Normative Systems as Spaces of Possibility

One way to understand the way games work is to see them as a system comprised of different types of rules that combine to create spaces of possibility and choice in which players have to pursue a specified goal or aim, under certain limitations. Taking soccer as an example, the aim of the game is to score goals, while the limitations players are under are that they aren't allowed to use their hands, or 'foul' their opponents, move off-side, etc. The rules will define what counts as a goal, what constitutes handball, off-side etc. So, the rules will consist of prescriptive imperatives that

7 Yudkowsky, Eliezer. 2008. "Artifiial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.

determine the aims or goals of the game, plus a raft of proscriptive rules laying out the limitations within which the players must pursue those aims, plus a number of definitional rules, or what Searle would call constitutive rules, which define what is to count as scoring, handball, off-side etc. Most games seem to conform to this general pattern of the pursuit of goals under certain limitations. It is easy to see this in games like Soccer, Cricket, Baseball etc, but it also applies to board games such as Chess, where players have a goal and limits on the way they are allowed to move each piece. The net effect is that the rules create an arena, or space of action, or space of possibilities, in which players are forced to make choices about how best to achieve the specified goal or aim, given the limitations under which they must play. In this sense the purpose of the rules seems to be not to direct players or to determine their actions, but to create new forms of choice and new dilemmas that didn't exist before the advent of the game. Even in games like Monopoly, where many of the player's actions are determined by imperative rules of the form 'if *x,* then do *y*', the structure of the game is such that it creates moments of dilemma and choice at specific points during the playing out of the game. Hence, games seem to be mechanisms for the creation of choices designed to extend and test the way we exercise our agency and creativity, and they achieve this through a carefully chosen set of rules which prescribe goals, define the terms of the game and proscribe the limits under which those goals are to be pursued.

One of the important things to notice about the space of possibilities that is opened up by any set of prescriptive and proscriptive rules is that it is, for all practical purposes, a space of infinite possibilities. It can accommodate an almost infinite number of possible strategies and tactics; there are an infinite number of ways to score a goal, a touchdown, a run, or to checkmate one's opponent, etc, so one can never exhaustively specify in advance all the possible tactics, moves or strategies that a rule space makes possible. The main reason one cannot exhaustively list all the possible strategies that future players might invent is because, as Ryle showed us[8], principles of strategy form a distinct set of rules which presuppose the rules of the game, but are not derivable from the rules of the game. There is no way that from my knowledge of the rules of chess I could logically deduce or otherwise predict that someone would one day invent the Sicilian Defence, or that the tactics of either bodyline bowling or sledging would be the inevitable outcome of the rules of cricket. So, if tactical principles are not logically deducible from the rules of the game, then the number of possible tactics any game may generate is limited only by the imagination, creativity and ingenuity of the players, and so, with enough ingenuity on the part of the players, there are, for all practical purposes, an infinite number of ways to interpret or apply the rules of most games.

From this we can see that games provide us with a perfect illustration of the interpretation problem in action. There are an infinite number of possible tactics, strategies or interpretations that stay within the rules of the game, just as there are an infinite number of ways an AI can achieve the goals we set it, no matter how many limitations we place on it. Furthermore, just as we saw with AIs, sometimes players in a game invent a new strategy that we consider undesirable, or don't approve of, despite its still being legally within the rules of the game. This is why the rules of games are continually evolving to take account of novel tactics that, for whatever reason, we think inappropriate or undesirable. But what happens when we make such modifications can be very instructive in revealing yet more complexity in the way rules systems work, because it forces us to ask why we find particular tactics undesirable and also what criteria we use as our guide when we alter the rules. What the following will hopefully demonstrate is that the rules of any game are formulated in relation to an external point or set of values that are not themselves represented in any rules of the game and may in fact be unrepresentable in principle.

The usual reason we modify the rules of a game is because the new tactic seems to clash somehow with what might be called the spirit or purpose or point of the game, so we alter the rules to outlaw it so as to maintain the spirit or values that the game is supposed to express. A good example of this occurred in American football in the first half of the twentieth century, when players invented a new strategy called the 'flying wedge'. A flying wedge was when the whole team linked arms to form a V-shaped wedge in front of the ball carrier and then charged headlong down the pitch as one. There was nothing in the rules that prohibited such a tactic and there was no way that anyone could have deduced or predicted from the then current rules of the game that this tactic would inevitably be utilised by someone. It was

8 Ryle.G, (2000), *The Concept of Mind,* London, Penguin. pp74-80

solely the product of the creativity and ingenuity of the players and it was extremely effective at scoring touchdowns if, as was usually the case, the opposing team could not find a way to break up the formation. It soon became clear, however, that this tactic had to be outlawed for a number of reasons. It made the game very dangerous, and also was a bit too effective and made the game more of a battle of brute strength than a game of artistry and skill. The tactic seemed to stifle the game and didn't encourage creativity and made the game boring both to play and to watch, and the fact that the flying wedge was outlawed shows that the main reason we invented the game in the first place was to encourage the development of just those things that the flying wedge stifled, e.g. creativity, flair, artistry, skill etc. And yet nowhere in the rules of American football are artistry, flair, skill, creativity etc mentioned. The game seems designed to promote certain values, yet nowhere in the rules are those values represented.

What this example demonstrates is that the rules of any game are formulated according to an external point or purpose. The rules of a game specify and define an internal goal, such as scoring touchdowns, but then we start to add proscriptive limitations which are designed to promote, encourage or otherwise generate certain types of activity, and it is the encouraging of these types of activities that is the external point or purpose of the game as a whole. The external point, purpose or values of the game dictate the rules we formulate, yet are not themselves represented in those rules. We invent all sorts of different games for all sorts of different reasons, to promote, encourage or create all sorts of different types of behaviours, and the behaviours we try to promote express what we value. However, we cannot predict from our knowledge of the rules what new strategies players will invent in the space of possibilities the rules have created, so we can never build a game that will guarantee players will not find creative new tactics and ways to achieve the internal goals of the game which subvert the values the game is supposed to express or promote. Consequently, the guardians of any game (i.e. whoever is in charge of deciding what the rules are in any given case) are in a continual arms race with the players of the game, just as we saw programmers would be in an arms race with artificial intelligences above. Players are always at liberty to use their creativity, imagination and skill to find new ways to interpret the rules of the game, and none of these new tactics or principles of strategy are deducible from the rules themselves, so they can't be predicted. Yet the guardians of the game have to continually ensure that the game expresses or promotes the sort of activities it was designed to promote and so continues to achieve the external point of the game, so they are continually playing catch-up to maintain the integrity and values the game is supposed to express.

The problem with all this is that not only are the external purpose or values of the game not represented in the rules, they cannot be, because we often don't know we value something until a player makes us realise that we do by inventing a strategy that we disapprove of. Often a new tactic will teach us things about what we value because it will undermine something we simply took for granted or hadn't thought was an issue, or hadn't realised we valued until it was undermined or lost, or the new tactic will undermine something we already knew we valued, but it will do so in a way we hadn't thought possible before. It wasn't until the flying wedge was invented that we realised how much we valued other aspects of the game that we had before taken for granted.

So, it seems that the values that drive our creation of the rules of a game are usually many in number, not explicitly stated in the rules, not explicitly formulated beforehand and are not even formulable beforehand.

In this sense it would seem that the values of a game form the assumed background against which it is created, but it is a background that is in principle never exhaustively articulable, because we can never specify in advance the various ways in which future players will create new tactics that undermine that which before was never under threat.

So it would seem that an agent who is playing any game is at the nexus of a very complex set of relationships between different types of rule system and unrepresented values: There are the rules of the game, comprised of prescriptive rules that determine the internal point or goals of the game and proscriptive rules which place limits on the ways players may pursue those internal goals; then there are principles of strategy, which are a distinct system of rules which are not derivable from the rules of the game and are potentially infinite in number and often take the form of prescriptive imperatives or rules of thumb, such as 'If *x,* then do *y*' ; then there is also the external point of the game or the values of the game, and these are nowhere represented in the rules of the game, but are what the rules of the game

are designed to promote or express by encouraging activity that conforms to those values. They are the unarticulated and inarticulable background of values against which we invent and continually modify the rules of the game.

All this seems uncontroversially to apply to games, but we wish to suggest that it applies to other, and perhaps even to all, normative systems such as the law and morality. Take for example the taxation system. Legislators are in a constant arms race with taxpayers (and their lawyers and accountants) who see the goal of the game as minimising their tax liability. They have a goal and certain limitations on the ways in which they are able legally to minimise their tax. This creates a space of possibilities in which they may exercise their creativity and ingenuity, and leads to them inventing ever more imaginative ways to avoid paying taxes within the rules of the game. Many of the ways they find, however, are clearly cases of following the letter rather than the spirit or values of the taxation system, so loopholes must be closed, but they can never all be predicted in advance. The criteria by which we judge whether a tactic is within the spirit or values of the game are such vague notions as 'fairness' and 'paying one's dues' and 'acting in good faith', and are never fully articulated or represented in tax law because they cannot be represented in a way that is not open to infinite interpretation, but they inform all our decisions about it. We cannot outlaw a tactic if we have not predicted it, and we cannot say beforehand what we value until someone does something that undermines something we previously took for granted.

Furthermore, we can see that the structure of rule systems outlined above is reflected in the structure of our legislative system more generally, where the legislature formulates rules and represents them in laws, but the judiciary is required to interpret those laws in practice in concrete cases. The judiciary's task is implement the spirit of the laws, since no law can adequately represent the values that grounded its formulation. What matters in the application of law is that injustice is not committed by means of the application of a law in the wrong way in particular cases. However, what counts as injustice can never be represented in an unambiguous or general rule because each case is context sensitive.

The question is whether the account of rule systems we find applicable to games and the tax system and other legal frameworks applies to morality and normative structures (or normative spaces) more generally. We suggest that it does. The existence of the interpretation problem in AI would imply that it does. In trying to programme morality into a machine in the form of proscriptive and prescriptive rules we encountered the problem that the machine is always capable of finding new ways to follow those rules, some of which ways we disapprove of. This would imply that any agent that is applying any system of proscriptive and prescriptive moral rules is operating in a space of moral possibilities opened up by those rules, and it is the fact that such a space would be one of infinite possibilities that generates the interpretation problem. Furthermore, the idea that morality can never be captured by determinate rules can be seen to be present from the very beginnings of western moral theory in the works of Plato and Aristotle. Plato argued that moral knowledge was abstract knowledge of the *forms,* which could not be represented as unambiguous rules[9]. Aristotle, while rejecting Plato's theory of forms, still agreed that moral knowledge could not be captured by rules, arguing instead that moral knowledge is a form of lived practice, skill or habit that could only be learned by habituation and cannot be captured in any rule book[10]. Moral knowledge, for Aristotle, was what modern philosophers, following Ryle, would call a case of *knowing-how,* rather than a case of *knowing that*[11]. Consequently, we suggest that the worry that moral knowledge is unrepresentable as rules and is somehow ineffable has always dogged moral philosophy. The interpretation problem in AI is just a vivid illustration of this disturbing possibility.

From our analysis of games, we saw that the values of the game are not represented in the rules of the game, and are not in fact representable at all because what we value forms the background against which we create the game and we cannot know ALL that we value because we cannot know ALL possible future tactics of players and how they might undermine values that had previously been taken for granted as part of the background. And if, as we are suggesting, normative systems all have this structure, then it would seem that the main reason we cannot programme our values into a machine is because our values are similarly unrepresentable in that they are the background against which we act

9 See, for example, Plato, Lee.D (trans) (1987) *The Republic,* Penguin, London (part 1 section 1)
10 See Aristotle, Thompson J. A. K. (trans) (1988), *The Nicomachean Ethics,* Penguin, London (Book two)
11 Ryle.G, (2000) *The Concept of Mind,* Penguin, London (Part 2 Section 3, pp28-32)

and found moral proscriptions and prescriptions, and we cannot know ALL that we value because we can never predict all possible future ways in which those background values might be undermined by the future activity of free agents acting in the space of infinite possibilities opened up by the rules we have in place.

If all this is correct, then it means that the interpretation problem is in principle unresolvable for normative systems in general. Not only do our background values guide us in creating the rules of any normative space, they are also what guide us in following those rules. Our creativity obviously makes it possible for us to follow the letter rather than the spirit of the rules if we choose to, but for the most part we are well aware of when we are doing this, and for the most part we attempt to interpret the rules in line with the values of the system, so the background values are a form of unrepresented knowledge that guides us and prevents us from interpreting the rules in unethical ways. So, the practical task that faces AI research is how to give a machine the background knowledge, or values that would prevent it from interpreting rules in unethical ways, if that background knowledge is always external to any rules system and is in principle unrepresentable. This is the task we wish to begin to address in the remainder of this essay.

Mistakes an AI could make - Mistakes of Intention and Instrumental Mistakes

To move forward with our analysis, we should try to get a clearer idea of the different ways an artificial agent could make mistakes, and, as mentioned above, we suggest there are at least two types of mistake it might make. To start with, consider that an artificial agent should have as part of its reasoning mechanism two major components, which can be further specialised, and not necessarily separated from each other. The first component will deal with what it ought to do, imperatives and obligations (given or inferred), and the second one will deal with the facts or beliefs it holds. This division is similar to that between the inference engine and the knowledge base in expert systems[12].

Upon trying to work out what the salient features are of the situations and of the nature of the mistakes in reasoning, we argue that the mistakes an artificial moral agent could make can be split into two categories corresponding to the two parts of the reasoning mechanism described above. These are mistakes of intention and instrumental mistakes, and we wish to argue that because there exist these two types of mistake, an AI must include an explicit moral program distinct from its practical reasoning program[13].

Mistakes of intention are mistakes about the imperatives or obligations that the agent has: about goals or limits. The mistakes about goals are occasions on which the agent errs with regard to the goals it holds, should hold, or has been given, whether in the task at hand or in general, whether given externally, by another agent, or internally, by its own reasoning. The mistakes about limits concern the errors that have to do with actions it should not perform.

For example, in the paperclip maximiser scenario, a mistake about goals could stem from the fact that the specified goal had no or inappropriate qualifications (no limits), and thus the agent might not know when to stop, or over what domain to maximise the number of paperclips, or it might use up an inappropriate amount of resources in pursuing the goal. Even when trying to specify the goal in detail using appropriate qualifications, the agent's creativity might lead it to make mistakes about goals. The agent could commit such a mistake when asked to "maximise X in this box using only Y kg of matter", by using matter of the wrong nature, like cooking the cat[14], or by using excessive amounts of power, as there was no explicit quantification for this.

Another example is that it could infer subgoals, such as having to continue existing for it to be able to achieve its goal, which might lead it to perform unwanted actions to stop itself from ceasing to exist. For example, in popular culture, Hal from "2001: A Space Odyssey" wants to avoid being turned off, killing humans in the process. The agents we build might formulate the sub-goal of having to stay on, absolutely, but we might not want them to stay on at all costs.

12 Jackson, P. (1998). Introduction to Expert Systems 3rd edition, p. 3. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

13 This point, that an explicit moral reasoning program is required, distinct from a practical reasoning program, has been made forcefully before by Wallach & Allen: See Wallach.W And Allen.C, (2009) *Moral Machines: Teaching Robots Right from Wrong,* OUP, New Yor*k*

14 Russell https://goo.gl/2EQ1zQ

Even if we tried to overcome mistakes of intention by trying to program into the agent some sort of moral framework based on a positive account, perhaps deontological or consequentialist, assuming we could, we would still encounter such mistakes, because there might be many circumstances we will not have accounted for while doing it, as it would be impossible to account for all possible circumstances. For example, a mistake about limits, to use the smile maximiser scenario, is harming humans by electrocuting their facial muscles in order to obtain forced “smiles”. The agent, perhaps, either did not correctly identify not harming humans as a categorical imperative (if it had a deontological framework), or did not recognise that not harming humans ought to be more desirable than obtaining more smiles (in a more consequentialist one).

Instrumental mistakes, the second type of mistakes an agent can make, refer to issues with the part of the agent that deals with facts or beliefs about the world (its knowledge base), similarly to the “failure of understanding” in Kantian terms. That is, when the agent does not correctly understand or predict how the world actually works, when it is making a factual or empirical error in its reasoning.

This type of mistake could occur due to a lack or failure of the agent’s common-sense knowledge, and its sources might be false beliefs, incorrect facts, or inappropriate/incomplete understanding of consequences of actions, amongst others.

An example is when the smile maximising agent does not correctly understand what a 'smile' is and goes on to tile the world with pictures of smiles, or paralyses people's faces into a rictus, because it mistakenly believes that they count as 'smiles'. Or, in broader terms, an instrumental mistake might be when a machine cannot apply the rule 'Don't harm humans' because its practical knowledge base does not allow it to grasp all the ways that humans can be harmed or how its actions might have harmful consequences.

Interestingly, we can see that the smile maximiser is making both kinds of mistake, instrumental and of intention. Electrocuting people’s faces into a ‘smile’ can illustrate a mistake of intention on the part of the agent, by misunderstanding what the given goal was or lacking understanding of its limits, or an instrumental mistake, by not understanding what a true 'smile' is.

Thus, even if we somehow overcame or minimised the instrumental mistakes, by improving its sensors, effectors or common-sense reasoning, the agent might still commit mistakes of intention, and this illustrates very well why we must attend not only to the agent’s practical understanding of the world, which will ward against instrumental mistakes, but must also build into its reasoning system an explicit moral component, to ward against mistakes of intention.

We can better see why this is by drawing on our previous analysis of rule following in games and normative systems more generally. The structure of a normative system together with the goals that are usually present in the rules of such a system and the fact that an artificial agent has similar rules in the context of practical reasoning in AI, allow us to see parallels between practical reasoning in AI (of which moral reasoning is a part) and human-centred normative systems, for example gameplaying or law. We can thus attempt to leverage the existing work on such normative systems, and in particular the above discussion on games, since agents acting in the context of decision problems in AI can be seen as agents acting in the context of a game.

When we want to implement an artificial moral agent in the context of any decision problem, we might require it have two types of goals: the purely practical, which are almost always explicitly provided through goals it has, and the moral, which are almost never explicit, having more to do with the external point, or values of the game than its rules. As we have seen, practical goals are immediately vulnerable to the interpretation problem, especially when given in the form 'do X' or another formulation that holds a specific desired behaviour: "maximise paperclips" for example. We have also seen that any finite limitations, or specific behaviours we place in the same form as part of the rules of the game are not enough to keep the agent on track with being moral, being themselves subject to the interpretation problem, so we need an explicit approach for moral behaviour.

We need explicit moral reasoning to be implemented in the agent if we want moral behaviour, because due to the interpretation problem we cannot rely on moral behaviour to come as a side effect of purely (non-moral) practical goal-

driven behaviour. An important reason for this is as follows. The practical, traditional, goals the AI has are part of the specifications of the decision problem, just like they are part of the rules of a game, but as in any game the player, or agent, can come up with ways of acting within the rules of the game but against the external point of the game, because this is not explicitly represented, so therefore we need another mechanism to keep the agent tethered to this external point.

To this end, we propose the use of active moral considerations, values, to inform the moral reasoning. To implement this, we could employ moral goals built around this type of values. To avoid the pitfalls of purely practical goals, these values should be *explicit* and *efficacious*, that is, be directly present in the agent's reasoning structures, and have a material impact upon the decision making of an agent in any relevant situations it acts in. We could then imagine having the agent prioritise or prefer these moral goals over the secondary, practical goals, in order to ensure that the former are not overruled by the latter. We will see more on values, and more details about such an approach, later.

Interpretative exposure law and causal power

As outlined above, we have seen that there is a lawlike relation between the amount of causal power a machine has and the degree to which we are exposed to the interpretation problem. There would be a qualitative difference between the consequences of behaviour by disembodied agents with purely digital causal power, and that of embodied fully autonomous agents with very capable sensors and effectors. This is because the former could only act digitally, perhaps communicating with other agents or controlling other digital systems, while the latter could directly impact upon the world and perform physical actions. The more restricted the causal power of the agent, the less unwanted effects it can have on the world, both quantitatively, as the domain it acts in is restricted, and qualitatively, as it can do less impactful actions with minimal (or no) sensors and effectors.

How can we use this insight in order to make progress? Firstly, we can consider the causal power we will assign an agent in the design process, and adjust it to the amount which we believe our reasoning mechanisms can successfully handle. For example, some could argue that perhaps a moral advisor is safest, at least to start with.

Secondly, we can split the reasoning into two parts to focus on one or the other. The two parts would mirror the types of mistakes an AI can make, with intentional mistakes covered by a moral (or practical) reasoning system and instrumental mistakes covered by an purely instrumental reasoning system. We might choose to do this only provisionally, in order to isolate one and start working on it. For example, we could decide we might want to work on the moral reasoning system first. To do this we could begin by providing the instrumental understanding required ourselves by, say, hardcoding it into the agent, thus eliminating any instrumental mistakes, and focus on examining its moral reasoning system, looking for and addressing mistakes of intention. This might prove to be especially useful as we do not yet have the sophisticated sensors and effectors required for either human-level perception, common sense reasoning or understanding of the world and this method would allow us to focus on the moral reasoning without having to deal with all that.

In what follows, we adopt this approach and focus on analysing and building the first part, the moral reasoning system.

Evaluation, building a 'character', and moral paradigms

To start with, we should ask ourselves what our end goal in this endeavour is, or more precisely, how we would tell that we have succeeded. How would we evaluate an artificial ethical agent**,** once we had it?

We might be tempted to require that the agent be able to justify its solutions to action selection problems by the use of an exhaustive enumeration of actions based on a causal account. To do this, we might want it to represent explicitly the reasoning behind the decisions it makes purely in the form of imperatives like “if S, do X then Y”, but we might not be able to accurately judge it by doing only this, if the interpretation problem holds. That is, perhaps there will be no satisfactory way for us to have the agent motivate its moral decisions by having it present only a conclusive end-to-end chain of imperatives or obligations to follow, imperatives in the form of “One is moral in a situation of type S if and only if one first performs action X, followed by action Y etc". This is even before we mention the difficulty of deciding upon

and including in advance, or alternatively learning at runtime, such a chain of imperatives for any possible variation upon the situation to be encountered.

Perhaps we will then need to proceed by getting the agent to give us some (any) reasoned answer to start with, and then evaluate and modify the reasoning system we have, either manually or automatically through a learning process, without expecting it to be able to identify in a representable way the solution to complicated ethical problems before this testing and training.

Regardless, we could require that it be able to *give a justification* for why it chose to act in a particular way, in order for us to gauge whether it is indeed moral or not. It could give an explanation based on the decision-making system it contains, the steps it follows in its reasoning (as we saw above) or, perhaps, the relevant values that informed its choice. Maybe it can show us the relevant values or considerations that it holds, and by looking at this and the agent's step-by-step reasoning and implementation, we can piece together how the moral decision making occurred and pass judgement on the whole package and the types of behaviours it exhibits, on this 'character' that it would have.

This is another reason why perhaps the key to building a moral agent that we can then successfully evaluate is that we build values in to inform its reasoning and make the above process of training and evaluation possible.

From a technical point of view this 'character' could be made of two parts, an interrelated mechanism of explicit values and considerations that we might call the *moral paradigm*, and a corresponding *moral reasoning engine* that handles their application.

However, to build this we first need to ask ourselves how we would expect or want the agent to act. An essential question that we are faced with is, then, what moral paradigm should we put in? As we have mentioned before, we cannot straightforwardly use classical versions of utilitarianism or deontology, due to the interpretation problem and their rule-based nature. So how, then, do we get it to understand the values that keep human rule following behaviour within acceptable moral bounds and what could we employ to encode that into it?

Ways forward - tackling the interpretation problem

To address these two questions, we use our work on rule following and Wittgenstein. In the vein of the later Wittgenstein, it could be argued that we need a different approach to explaining philosophical concepts, and moral paradigms more specifically, than saying what rules they might consist of or laying down a direct representation of their content. This is because this approach might be impossible, if conveying meaning through a representational medium is subject to the interpretation problem. If we are indeed faced with this issue of representation, then we require another method to employ when transmitting meaning to the artificial agent, so as to mitigate the effects of the interpretation problem as much as possible.

A glimpse of one answer to this can be seen in the "Investigations", which is itself a work that attempts to take us on a journey to the meaning behind Wittgenstein's words, itself arguing through its structure that there is no representation, in language, that can accurately convey that meaning. It is this concept of showing, rather than saying, that could inform our approach as well.

But how could we show the meaning to the agent, if that is what we wish to attempt? Perhaps we can take a page from Wittgenstein's book and try to get the agent to understand our meaning by taking it on a journey (going through a process). That is to say, by giving it a device which can help it indirectly to understand what we want it to do and how we want it to be.

In the case of moral behaviour and practical reasoning, more generally, we argue that we can use the concept of *values* to do this. By values we mean high-level concepts that are relevant considerations during decision making. These could be virtues, character traits ("honesty" for example), moral concepts that are of importance to the agent or others ("property" for example), or general (morally neutral) practical considerations (to deal with purely non-moral practical reasoning). The essential aspect to them is that they are the tether to the external point of the game, the concepts that

crystallise what we want from the game being played, or the moral situation being addressed, from the behaviour of the agents. We believe that keeping the concept of 'value' as general as possible will help both in order to hedge against the interpretation problem, as exemplified below, and in order to give a wide spectrum of application for any value based framework developed from our work. For instance, one might wish to ignore all our focus on morality and still use such a framework for building purely practical goal oriented artificial agents.

An important reason behind the usefulness of values is precisely this ambiguity, and the fact that they are multiply realisable. That is to say, they can be embodied, or promoted, by different actions (even in the same context), and can be adhered to by a plethora of behaviours. Most importantly, the intrinsic flexibility that the concept of 'value' has helps us hedge the risk that the interpretation problem presents.

But if we wish to keep values as ambiguous as possible, then, it could be asked, aren't we giving up the quest of getting an agent to do what we want, if we give it leeway to perform differently to a specific desired behaviour? Wasn't getting the agent to do exactly what we wanted the whole point of building our AI in the first place?

To this, we would say that yes, our goal in building AI is obtaining some type of behaviour from it, but what we want from an even minimally independent practical agent is not for it to do action X, followed by Y, and then stop, but rather to achieve such and such a complex goal, yet acting in such and such a way while doing it. The purpose of a value based approach would be to allow the agent freedom to improvise in practical terms, while ensuring that it exhibited a certain character while doing so

The whole point behind the interpretation problem is that, in a sufficiently complex environment (and certainly in any sort of environment that aims to simulate the real world or parts of it) any goal or behaviour or piece of meaning, when conveyed in a representational way (using language, written down or otherwise codified), needs to be interpreted and therefore can be misinterpreted and misunderstood. Therefore, what we want from the AI is always a complex construct, due to, amongst other factors, the nature of language itself. Similarly, any other representational tool, besides language, will fall prey to this due to interpretation still being required, and it will always be vulnerable to the interpretation problem.

If we recognise this, then we can move forward by attempting to convey the meaning that we desire using more abstract constructs, namely values, attempting to show indirectly, rather than tell. To start with, some values might be amenable to definitions for a particular purpose, and they could form anchors (for example 'property' and its definitions in law). The idea here is that there can be some concrete starting point in terms of programming values for this process of building a moral agent, and we then need a way to get it to act in the spirit of these values. This we could do by giving the agent the goal of solving the problem of becoming a certain type of character, by building an explicit value-based moral paradigm into its reasoning, and then examining the results we get and iteratively fine-tuning it (manually or automatically) based on its behaviour. In this way, the machine would be using its creative ability to maximise the primary goal of, say, being trusted by its trainers, or being considered honest and so forth, so we would be getting the machine to do some of the work of solving the interpretation problem for us by giving it the goal of exhibiting certain virtues.

On the implementation of a value-based approach

Below, we propose some areas to examine and issues to consider while building artificial moral agents. We look at the structure of the value system, the nature of the values used, an idea for testing and training, designing normative systems for artificial agents and some important aspects of 'the Background' against which agents act in a normative system, to do with the previous knowledge or goals and embodiment of the agents.

To move forward with a value-based approach as described above, an aspect we should pay attention to is the structure of the value system the agent has. Just having values represented explicitly in the reasoning might be beneficial in itself, but to further add meaning to the system we can leverage the structure itself and the relative placement and interactions between the values in the moral paradigm or character of the agent, as well.

A structured approach, wherein we would specifically choose a structure such that it furthers our desired use of the values, seems essential, as we do after all need a representable way of building a reasoning system for AI, as any kind of programming for it is a representation, and due to the interpretation problem, this representation cannot, with perfect precision, come in the form of purely practical imperatives or general rules. So it might not be enough just to 'have' values. The structure must provide the interaction between values, the glue that holds them together and the arena in which they can perform together. An example is using a preference relation to structure the values, wherein one would classify them based on their relative importance using some order structure. In a forthcoming work, we demonstrate a hierarchical preference ordering for building moral paradigms, based on a qualitative difference between the strata.

Another way of adding accuracy to the meaning of the solution is for it to come from the type of values that we use. We have mentioned before that, for example, the values could come in the form of virtues, but perhaps instead of having a set of simple values, such as 'Honesty', we might use a set of more complex, human-dependent, values, such as 'Being trusted'. We could call such values 'relational values', because they are intrinsically relative to another agent, for example 'being trusted' is relative to a particular human (or any human) and its opinion of the agent. The nature of these relational values could shift the problem of understanding posed by the interpretation problem, because now the agent does not need to understand what the values are in isolation, by itself, with no quantifiable evaluation of success, but rather can only understand them in relation to us. This makes it easier for it to understand what we want it to do, as it has a tangible source for its values, and for us to understand and evaluate what it does, as the values are directly based on us and thus we can more easily evaluate their status.

Training

From a technical point of view of how one might get the agent to understand the values and succeeding at promoting them, and testing it, one can imagine, for example, a training mechanism by which the machine would gain understanding of what the values mean or how they are quantified, how they are achieved, by interacting with human agents, with the humans providing the measurement of its success, either by reinforcement learning, or manually modifying the algorithm after observing its decisions etc. Using reinforcement learning, for example, has been suggested before [15] and using this type of training process for the goal of specifically learning human values has been attempted in [16] and [17]. But the crucial difference between these approaches and ours lies in the type of values that are to be learned, as we discuss below, and approaching the whole endeavour not as a value specification problem, aiming at some representational content that is to suggest understanding on the part of the agent, but a character-building one, aiming at a non-representational type of behaviour that is to ensure the desired behaviour. This would hedge against the interpretation problem, which as we have argued would be ever-present and unavoidable in a purely representational approach.

For example, using the relational values discussed above, the agent might act or give a solution to a decision problem, based upon its desire to, say, become 'trusted' (having that as a relational value), and we could then collect feedback from a tester or trainer (or trainers) of whether it is now more or less trusted than before, or whether the action it has selected manifests the value of 'being trusted', and make the appropriate evaluation of its reasoning system and the required modifications. This might feedback through an automated learning process, or indirectly by our tweaking its reasoning system based on our judgement of its actions or decisions.

The “Background”

With regards to future work, there are some very important starting points that we can examine and use in order to make progress. All aspects that have to do with the previous experience and goals of the agent, the games it plays and how these games might interact, together with details of its embodiment, fall into what we might call the "Background".

15 Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). Alignment for advanced machine learning systems p.5. *Machine Intelligence Research Institute*.
16 Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. (2016). “Cooperative Inverse Reinforcement Learning.” arXiv: 1606.03137 [cs.AI].
17 Everitt, Tom, and Marcus Hutter. (2016). “Avoiding wireheading with value reinforcement learning.” arXiv: 1605.03143.

These have to do with the nature of the agent, and the context in which the agent acts from in any game, and we should try to leverage or at least consider this Background. To do this, we might think about, for example, the pre-existing goals, knowledge, or aims the agent might have, its physical capabilities (what sensors or effectors it has as part of its embodiment), its linguistic abilities (how it might interpret the wording of the rules) or its previous experience with similar rules or normative systems.

Conclusion - On Value Alignment

To conclude, what we have argued throughout this paper is that there is an important distinction between rules and values. Moral rules, in the form of practical imperatives for example, correspond to the rules of the game, while the values correspond to the spirit of the game, the external point of the game. These values can only be shown, not said in the sense of being accurately transmitted through a representational medium. The tension that arises between following the rules of the game and the spirit of the game, between the moral rules we give the artificial agent and the kind of behaviour we want it to achieve, gives rise to the Interpretation Problem.

For this reason, as well as others, the process of value alignment is very important for obtaining moral behaviour in agents. However, most attempts to achieve value alignment look at implementing the same rules we seem to follow into the machine [18] [19]. But just as we can misinterpret the rules, so could a machine, and thus perhaps we should instead focus on aligning the interpretation of the rules. Since it is the spirit of the rules that needs to be understood and acted upon, this, we have suggested, could be done through a value-based or character-based approach, using the agent's power of reaching goals to get it to aim at becoming a certain type of character of which we can approve.

References

1. Mariëtte Le Roux (12 March 2016). "Rise of the Machines: Keep an eye on AI, experts warn". Phys.org. Retrieved 13 March 2016.
2. Russell, Stuart J.; Norvig, Peter (2003), Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-790395-2. pp. 27, 32–58, 968–972

18 Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). Alignment for advanced machine learning systems p.5. *Machine Intelligence Research Institute*.
19 Soares, Nate. (2016). "The Value Learning Problem." In Ethics for Artificial Intelligence Workshop at IJCAI-16. New York, NY.