Strawson's take on moral responsibility applied to Intelligent Systems

Laura Cromzigt
3739953

7,5 ECTS Bachelor Thesis Artificial Intelligence

Utrecht University

8th January 2015

Abstract

This essay investigates the attribution of responsibility to intelligent systems. It argues that more traditional approaches to the subject, mainly independent conditions theories, fail because they encounter some of the same problems that one encounters in attributing responsibility to human agents. A very different approach, namely Peter Strawson's take on responsibility, is introduced and applied to intelligent systems. Strawson claims that theoretical considerations miss the point when we ponder the responsibility of human agents. He claims that we should understand assigning responsibility as part of the practice of human life. This claim is investigated and transferred to intelligent systems to see if it provides a more fruitful way to understand responsibility of intelligent systems.

Keywords Strawson, responsibility, Independent Conditions Theories, Artificial Intelligence, Intelligent Systems, moral community

Contents

Introduction

As Artificial Intelligence becomes more and more advanced questions about responsibility arise. For example, who is responsible when artificial intelligent systems (IS[1]) cause an accident? Who is to blame when a driverless car fails to brake and hits someone? Who should we accuse if a drone inadvertently kills innocent school children? IS, such as drones and driverless cars, are considered learning (i.e. semi-autonomous)[2] automata and are thus not merely tools in the hands of human agents, fully controlled by the operator [Matthias, 2004][3]. As IS become more and more part of our human world and potential harm issuing from their actions is no longer a theme for science-fiction writers, we should think about who (or what?) to blame when things go wrong.[4]

Questions surrounding IS and responsibility are or seem relatively new compared to questions on human agents and their responsibility. For thousands of years we have discussed when and how to ascribe responsibility to whom and why. The arguments used in these discussions have had time to develop and it may be that we can use them to understand when and how to ascribe responsibility to IS. Therefore, in this essay the most common way responsibility is attributed to humans, namely by checking to see if so-called independent conditions are satisfied, is investigated.

[1] In this essay the abbreviation IS is sometimes used for the plural, sometimes for the singular.

[2] Adaptive software enables a formal system to adjust its reactions to variable stimuli. Such a system displays behaviour that is informed by experience, that has adjusted its input-output relations based on a history of trial and error. Such a system displays behaviour that is the result of learning. Whether such a system acts autonomously is a question that cannot be answered here. Such systems are called semi-autonomous to stress that their actions cannot be traced back to the actions of designers, engineers, programmers, etc. In section 3 the question concerning autonomy is discussed in more detail.

[3] The gap between the delegation of responsibility and the assignment of responsibility in intelligent systems is described by Andreas Matthias as follows: "we will see how certain recent developments in the way of manufacturing computerised, highly adaptive, autonomously operating devices, inevitably lead to a partial loss of the operators control over the device. At the same time, the degree in which our society depends on the use of such devices is increasing fast, and it seems unlikely that we will be able or willing to abstain from their use in the future. Thus, we face an ever-widening *responsibility gap*". [Matthias, 2004, p. 176]

[4] Recent developments in the field of Artificial Intelligence, with IS becoming more advanced, have led to a renewed interest in responsibility and IS. Broersen describes that we delegate more and more responsibility to intelligent devices, ranging from self-driving cars to autonomous trading systems. Only where things have gone wrong the problem of responsibility is recognized as such [Broersen, 2014].

The application of independent conditions is usually considered fraught with problems on a purely theoretical level, but intuitively possible on a practical level [Braham and Van Hees, 2012]. We tend to know when some human agent can be blamed for a certain action, and most of the time we know when blame is not applicable. There are not that many situations in everyday life in which the theoretical basis of attributing responsibility has to come to the fore and explain itself[5]. That is not to say that the attribution of responsibility just outside the realm of everyday life is straightforward.

In Section 2.1 three conditions for ascribing responsibility are discussed. These three conditions are 1) autonomy, 2) having reasonable alternatives and 3) being causally relevant. Autonomy sometimes gets discussed when we are concerned with developing children and the sanity of grown-ups. Whether someone had reasonable alternatives for his[6] actions depends on the definition of reasonableness. This definition sometimes gets discussed in real life when trying to discern between temptation and compulsion. Whether ones action is causally relevant to some event becomes interesting when discussing global causal connections, such as ones contribution to global warming.

The advent of IS in our world, however, presents us with a whole new set of (possible) situations that elicit our doubt. IS are undoubtedly agents in that they act and that their actions have consequences for us humans[7]. Some of the acts IS are capable of would normally cause us to search for responsible agents to blame or praise. We ask questions like 'Who did this?' and 'Who can we reprimand or reward?'. In the case where an IS (semi-)autonomously acted it becomes difficult to find this responsible agent. Neither the designer, nor the operator can be reprimanded or rewarded: the IS itself now has to be questioned [Matthias, 2004].

It will be found that, unfortunately, the application of independent conditions to human agents fails on a practical level. The failure can be seen as stemming from an incomplete understanding of the conditions themselves. It will be argued that on a philosophical level it cannot be ascertained whether these conditions apply. Not even in the case of a human act, let alone that

[5]Although developments in neurology raise questions on responsibility and accountability (for instance in the area of law, see `http://www.njb.nl/Uploads/Magazine/PDF/NJB-1345.pdf`), mostly focussing on some account of free will, questioning whether someone can and should be held responsible for some act remains a matter of exception, not custom.

[6]In this essay I am using 'his' and 'he' for brevity, but it should be read as 'his or her' and 'he or she'.

[7]Section 1.1 below gives a full analysis of what it means to claim agency for IS.

such independent condition theories can help us in understanding responsibility in IS.

Are we then left empty-handed? Do we have to develop a firm understanding of what it means to apply the theoretical concepts to human actions before we can decide on the applicability of responsibility to IS? Is the situation in which we hesitate to attribute responsibility to human agents the key? Should we withhold judgement until the conceptual problems are resolved, or should we look at the situations in which we intuitively know when an agent is to be held responsible? In this essay I will use Peter Strawson's ideas to claim that we do not have to withhold judgement by analysing why and how we normally attribute responsibility.

In the first part of this essay, section 1, it is argued that IS are or soon will be autonomous to the point where no designer, operator or other human agent can predict, fully explain or control the actions of IS, much like our knowledge of or control over the actions of a human agent is limited. In the second section the more traditional way of evaluating responsibility is sketched (section 2.1) and applied to IS (section 2.2). In the third section Strawson's ideas on assigning responsibility are investigated (section 3) and applied to IS (section 3.2). In the fourth section I will give six preliminary observations for future research (section 4.1.) and statements on the possible moral membership of IS will be discussed (section 4.2.).

1 Intelligent Systems as agents

1.1 Intelligent Systems are or will be autonomous agents

What is an intelligent system? Intelligence is notoriously hard to define, but in the context of the main theme of this essay (assigning responsibility to IS) intelligence is understood as the disposition that enables an agent, human or other, to autonomously decide on what to do, based on reasons it can have and weigh. This means that the actions of an intelligent agent are by definition not totally controlled by other agents, since the agent would then not be autonomous. The actions of the agent cannot be understood as random, but must be understood as part of a plan, a goal or an intention to act and knowingly cause consequences.

It is this interpretation of intelligence this essay uses when discussing IS. This definition seems to make the main question (how to and how not to assign responsibility to IS) circular. I have defined being an agent as being able to act intelligently, and have defined responsible as having acted intelligently; then we can assume that all agents are appropriate targets for assigning responsibility. If we define IS as systems which are free in the sense of being autonomous and intentional, why did the issue of attributing or withholding responsibility arise? If their capacities in this regard are the same as ours, and we, humans, are suitable targets for assigning blame or praise, why do we hesitate to assign blame or praise to IS? The main issue is whether IS are indeed intelligent agents, and thus might be blameworthy or praiseworthy. In this section IS are used to designate systems (that is, coherent compositions that have one or more describable functions) that act as if they are intelligent, as if they have intentions and autonomy.

An IS is an adaptive system, a system which learns to adapt its actions on the basis of feedback from the environment in which previous actions have taken place and in which future actions must take place. The adaptiveness from interaction with the environment is an important trait in IS, since the variables which a system encounters are unpredictable. When the system adapts its output in answer to this varying environment the output itself becomes unpredictable. The system has, of course, an assigned function given by his or her designer, but this function can be quite diffuse for complex systems. "See to it that parts x and y are welded" is an example of a non-complex function because the margins of action for the system are narrow. A system which is designed for such a narrow function will not display very unpredictable behaviour. "See to it that no suspect person enters this building unchecked" is an example of a diffuse (or wide) function

assigned to highly complex systems. Such systems will have to deal with an enormous amount of variables, will have to learn from their actions and the reactions, and will display lots of unpredictable behaviour. It is not the result of such systems being highly complicated that makes their actions inscrutable, it is a result of their being adaptive to an inherently unpredictable environment. In this case no engineer can predict with certainty what an IS will do because no engineer can predict with certainty what the environment will do. As explained earlier, adaptive (dynamic) software enables the artificial system to learn to adjust to the challenges the environment throws up. The system, of course, needs some sort of mission (very like humans need some sort of motivation) to give it its goal-oriented definitions of success and failure, without which its trial and error strategy would be meaningless. A non-intelligent system is a system which is not adaptive or a system which is unable to modify its behaviour because of certain constraints (lack of computational power, lack of motoric adjustability, etc). An IS can and will adjust to its environment and adjust its own actions (and possibly adjust the environment) to match its goals. Between non-intelligent and highly intelligent there are, of course, many different intermediate stages of intelligence. These can all be defined by checking how well they can manipulate themselves and their environment in order to achieve some goal.

Although there is no clear demarcation between non-intelligent systems and IS we can safely claim that we are already and will, in vastly greater numbers, be surrounded by IS. Google says it will put a driverless car on the roads by 2018[8], drone-acts are closely resembling self-coordinating attitudes[9], your fridge may be planning your dinner[10] and your pc is crammed with all kinds of bots that filter your searches based on what they have learned about your interests and where you will most likely spend your money. Most of these systems go unnoticed up until the point where they cause some kind of problem. In those situations we curse our machines and usually remember that they are just machines, without intention to frustrate us or even harm us. In the case of IS, however, we sometimes hesitate: is a system that has a diffuse task adaptable enough to have malevolent intentions?[11] HAL from the Kubrick movie '2001, A Space Odyssey' comes

[8] `http://www.bbc.com/news/technology-27587558`

[9] `http://www.nature.com/news/autonomous-drones-flock-like-birds-1.14776`

[10] `http://www.usatoday.com/story/money/business/2013/04/08/high-tech-home-improvement-kitchens/2043807/`

[11] An unanswered question arises: is the task diffuse for the machine itself, or does it simply seem diffuse because the task is too complicated for us to comprehend? The answer to this relies on our understanding of the IS: are we willing to assign it some kind

to mind. Its task is to protect the mission, not the people involved in the mission and can thus be able to harm intentionally [Dennett, 1997]. It is at this time the autonomy and intentionality of IS become cause for serious ethical scrutiny: must we, can we and if so, how do we determine if we can blame or praise an IS? Is an IS an agent suitable for ethical appraisal? To understand this question we must begin to understand what an ethical agent is.

1.2 Ethical agents

[Moor, 2006] discerns three kinds of ethical agents:

implicit ethical agent An implicit ethical agent is an agent who has a strictly defined set of actions. Examples of implicit ethical agents are Automated Teller Machines. An implicit ethical agent does not have to think[12] about what is right and it does not have to weigh reasons. The system is defined by a narrow function and has little to do with the environment: there are only a limited number of possible actions and the environment in all its diversity and unpredictability is reduced to 5 or 10 different states, depending on how many buttons and slots the ATM has. These kind of IS (if they are intelligent) are common in our human worlds and fit the traditional or laymans view of what a computer is, namely a device in which all actions are determined by algorithms the designer has installed. The system may be very complex, but the designer knows or can calculate what the output will be given a certain input.

explicit ethical agents Explicit ethical agents are agents who use an ethical model given by their designer to make choices for their actions. For example, an IS programmed with a Hedonistic Act Utilitarian model will make choices based on something like a hedonistic version of the greatest happiness principle. Or a Kantian line may be installed, wherein human life has to be spared at all cost. Drones for example can be equipped with a certain amount of autonomy. There are scenarios in which a drone can decide and act for itself, without human intervention. These decisions can be informed by a narrow

of consciousness so it can fathom its own tasks or are we limiting the IS to the role of executor of whatever task we deem it fit? This question is beyond the scope of this essay although a related question is raised in section 2.2.2.

[12]When I use the word 'think' it is not implied that IS can really think like humans. That question is irrelevant in the context of this essay.

assignment, e.g. see to it that you fly at this altitude, or by a wide assignment, e.g. see to it that no suspect person enters this building unchecked. In this latter assignment the programmers or controllers will have to have made certain decisions, they had to install certain criteria on which the drone can base its acts. These criteria, among which will be components of an ethical model, are then coded into the system as an unalterable basis. The drone cannot decide to change these criteria or ignore them[13].

full ethical agents A full ethical agent can make explicit ethical judgments and generally is competent to reasonably justify them. John Martin Fischer calls this ability to use reason to justify one's choices crucial. If an agent is able to give reasons for his choice we might assign him free will [Fischer, 1998]. To this moment, no IS is seriously considered a full ethical agent.

It is not unlikely that we will have an IS in the future which can make explicit ethical judgments like a full ethical agent. This essay will focus on IS as explicit ethical agents in Moor's sense. The considerations on IS and responsibility will, however, a fortiori apply to full ethical agents since the arguments in this essay use a sense of autonomy-assignment by members of a moral community to assign responsibility, and autonomy will more readily be assigned the closer an IS comes to being a full ethical agent.

[13]The most famous of such robot-laws are fiction writer Isaac Asimov's 'three laws of robotics', taken from his *I, Robot* [Asimov, 2004]

2 Attributing responsibility

2.1 Independent Conditions Theories

One widely held way to think about how to assign responsibility is to formulate some independent theoretical conditions on being responsible. According to this view, which I will call independent condition theories (ICT), these conditions have to be met if an agent (human or non-human) is to be held responsible. Most of the time, these conditions are held to be evident, intuitively true no matter how you look at them. It is somewhat of a truism that applying these conditions to what we consider to be normal circumstances or normal agents yields a nice fit. An action of a normal agent in normal circumstances will satisfy such conditions, simply because we have defined normal as fitting such conditions. The commonly proposed conditions will be sketched first via two examples of such theories before turning to the applicability to agents or circumstances that are not normal, like IS.

[Noorman, 2014] says that most philosophers think that the following three conditions are the independently necessary conditions, meaning that each is a necessary but not sufficient reason to assign responsibility for moral responsibility. I will call these sets of conditions Independent Conditions:

1. There should be a causal connection between the person and the outcome of actions. A person is usually only held responsible if he had some control over the outcome of events.

2. The subject has to have knowledge of and be able to consider the possible consequences of his actions. We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.

3. The subject has to be able to freely choose to act in a certain way. That is, it does not make sense to hold someone responsible for a harmful event if his actions were completely determined by outside forces.

Another example of an independent condition theory is Braham and Van Hees' slighty different version of ICT [Braham and Van Hees, 2012, p. 605]:

Agency Condition (AC) The person is an autonomous agent who performed his or her action intentionally.

Causal Relevancy Condition (CRC) There should be a causal relation between the action of the agent and the resultant state of affairs.

Avoidance Opportunity Condition (AOC) The agent should have had a reasonable opportunity to have done otherwise.

Normally, these (examples of) conditions are easy to apply, because an agent who freely and knowingly chooses from amongst several opportunities to do something that causes something is held responsible for acting. If I throw a brick and the brick hits someone then I am responsible. If I accidentally hurt someone then I am not responsible[14] (my act has not met the condition of being intentional, AC). If I fall from a window and hurt someone, I am not responsible (my act has not met the condition that I could have done otherwise, AOC). If I take a left turn and to my right someone falls over, I am not responsible (there is no causal connection, CRC[15]). The ICT become more of a challenge when we try to apply them to less normal situations, e.g. what if we doubt whether the agent acted intentionally? What if we doubt whether the agent had reasonable alternatives? Such situations are, in our human world, encountered when dealing with agents who are structurally or temporarily unable to reason. The severely mentally incapacitated are structurally unable to reason about their actions. Children are examples of agents who – normally – are temporarily unable to reason about their choices. These agents, at least at the moment when they acted, do not meet Noorman's condition number 2 or Braham and Van Hees' AC. There are also situations in which an agent could not reasonably have done otherwise. In these cases the act does not meet Noorman's condition number 3 or Braham and Van Hees' AOC. In section 2.4 it will be argued that from a certain perspective it is impossible to know whether these demands are met and that poses a problem for assigning responsibility to IS as well as to human agents.

2.2 Independent Conditions applied to Intelligent Systems

In the preceding section the most common of ICT, applied to human agents, were briefly sketched. In this section ICT will be applied to IS to see if this enables the attribution of responsibility to such systems. Braham and Van Hees' version of ICT will be used. It needs to be checked if we can verify all three conditions. They are all necessary, but neither one sufficient, so all three will have to apply.

[14] I might be liable, but this juridical concept is not relevant here.

[15] In this essay I will not question the causal relevancy, although questions like the many-hands problem could pose a problem here.

2.2.1 Causal Relevancy Condition

The event for which responsibility needs to be assigned had to have an act performed by the IS as a necessary condition. If the act was a necessary condition to the attributable event, the act satisfies this condition, CRC. This connection is considered empirically verifiable and not contested in this essay.

2.2.2 Agency Condition

In section 1.1 it is described what it means when we call an IS an agent. Now we must see if such an agent satisfies this condition, AC. An IS satisfies this condition if it performed the act intentionally, i.e. knowingly and willingly. This means that the IS has to have knowledge of what it is doing and what the consequences of doing so will be, and that it has some reasons to do what it does[16]. How can we tell? Does an IS who acts as if it knows what it's doing and who gives reasons for what it's doing really knows and really have reasons? In the lively debate between strong and weak AI such questions are crucial. Proponents of IS as truly intelligent systems claim that acting as if it is reasonable is reason enough to attribute reasonableness to a system, opponents claim that such acting is nothing but a hollow shell, an imitation of true reasonableness lacking some vital ingredient to be rightfully called reasonable[17]. There is as yet no clear answer to these questions. As long as these questions stay unanswered we do not know whether an IS meets the AC.

2.2.3 Avoidance Opportunity Condition

Finally, it needs to be checked if the act of an IS can satisfy the AOC. This means that the IS has to have had reasonable opportunity to do something else than it did. Most people consider IS such as computers as completely defined by their algorithms, determined in their actions and thus unable to do something other than their programming tells them to do. In section 1.1 it is argued that intelligent systems are not restricted in this way anymore

[16]The claim that it is impossible to know and control the consequences of one's actions for certain was a central argument in Kant's dismissal of utilitarian ethics. He argued that is it unreasonable to hold an agent accountable for the consequences of his actions since the agent does not control those consequences. Thomas Nagel has also explored this claim in his Moral Luck [Nagel, 1991, pp. 24-38]

[17]The most famous proponent is [Dennett, 1993], the most famous opponents are [Searle, 1980] and [Chalmers, 1995]. The key thought experiment is the Chinese Room experiment.

and can act in ways that are not predictable. They react to an unpredictable environment and adjust their actions based on their goals (efficiency being the main independent goal of any set or subset of missions) and the feedback they receive. Since IS base their operations not on necessary and sufficient instructions but solely on necessary instructions combined with goals, they adjust, learn and respond in a way like humans do. The discussion started in the above attempt to apply the AC to IS can be repeated and expanded here. IS can or soon will act like humans can act. The question whether there is something that, as some kind of underlying aspect, affects these actions in such a way that human actions are significantly different from IS actions, remains unanswered.

2.3 ICT not sufficient for ascribing responsibility to IS

The ICT way of thinking about responsibility is not sufficient when reasoning about responsibility of an IS because 1) it cannot be ascertained whether an IS meets the Agency Condition and 2) all the conditions need to be met for ascribing responsibility. Neither can it be ascertained that an IS meets the Avoidance Opportunity Condition. It might be conceded that present-day IS are obviously not even close to resembling human agents, and that their actions are therefore very different from human actions, but that is not what is at stake here. As will be argued in section 2.4, we cannot fully describe what it means to claim that a human act satisfies the AC or the AOC. It can only be claimed that human actions satisfy the CRC, something which can also be claimed for the role of IS in causing their acts.

Present-day and future IS come ever closer to acting like humans act. Claims about human-acts satisfying conditions which IS-acts do not satisfy are moot. In section 2.4 the philosophical implications of these problems are discussed. This leaves us with an unsatisfying conception of responsibility for humans and non-humans alike.

It is clear from these findings that we need to consider new approaches for thinking about assigning responsibility to IS. In the remaining part of this essay a first attempt will be made and an alternative way of understanding responsibility, Peter Strawson's, is introduced and applied to IS.

2.4 Problems of applying ICT to IS

In the previous section it had to be conceded that moral responsibility cannot be attributed to IS when using a kind of ICT. It is thus not clear what the actions of IS have to meet for IS to be held responsible for them. The Causal Relevancy Condition or the demand for a causal connection between the agent and the outcome of actions can be understood and verified, but the demand that an agent has 'to be able to freely choose' and has to 'have knowledge of and be able to consider the possible consequences of her actions' before being a proper moral agent is intuitively understood at best. This presents researchers concerned with the gap between IS and responsibility with a problem. As long as we cannot fully understand what 'acting intentionally' (Braham and Van Hees' AC) and 'having a reasonable opportunity to do otherwise' mean, responsibility cannot be attributed to IS. We simply cannot say anything on this subject.

This is, however, a problem that concerns more than the IS-responsibility gap. The problem we encountered in the foregoing section spells serious trouble for attributing responsibility to human agents as well. As with the actions of IS it is quite easy to understand the demand for a clear causal connection between an agent, an action and a consequence. As noticed, this condition is a necessary but not a sufficient condition for attributing responsibility. The agent has to have had the intention (i.e. has to act knowingly) and has to have a form of free will (the ability to act otherwise). Neither of these last two demands could unequivocally be applied to IS, but neither can they unequivocally be applied to humans. Humans, it is assumed, have consciousness (or are at least able to know what they are doing) and have, in morally relevant circumstances, a choice in what they are doing. However, these assumptions have been challenged and a number of arguments have been presented to refute them. These arguments will be briefly discussed.

The claim that humans are able to know what they are doing rests on the premise that humans have a mind that consists, among other faculties, in the ability to deliberate reasons, motives, possible actions and consequences of those actions. A human can know what the options are and can choose among them, based on reasons. Therefore some acts can be deliberate acts and the agent performing such acts can be said to have acted deliberately, intentionally. Only then can an agent be held responsible for his actions. An agent who unintentionally causes something cannot be held responsible. A young child that by playing sets of a series of unfortunate events is not held responsible in the sense an adult is held responsible. The adult is supposed

to be able to know what would happen if he acts in a certain way and the child is not supposed to know (yet)[18].

The problem with the premise that people have minds capable of such deliberation is that it is astoundingly difficult to prove or disprove that people have minds. This problem is known in the philosophy of mind as the *Problem of other Minds* and has given rise to theories that try to do away with the whole concept of mind, such as certain types of physicalism, reductive behaviourism and (early) functionalism. It is one thing to say that people are composed of such and such materials and that they act in certain ways, but it is quite another thing to claim that humans somehow perform or act intentionally[19]. This is not the place to fully explicate the arguments that are used by proponents and opponents of the intentionality of humans (and, possibly, higher sentient beings), but the gist is that it cannot conclusively be proven that humans have a faculty that makes - some of - their actions deliberate or intentional.

The second assumption underlying the independent conditions used in evaluating responsibility is the assumption that humans have free will. The following, somewhat limited[20], concept of free will is at stake here. An agent had to have a reasonable alternative to the action he chose to perform, otherwise he did not act out of free will and cannot be held responsible for his action. This version of freedom of the will is threatened by several kinds of determinism, most problematic being the mechanical determinism of Laplace's Demon. Determinism is a thesis that holds that everything that happens is caused solely by events in the past, and everything that will happen is caused by the present events. That leaves out any chance of things happening any other way than the mechanical laws of cause and effect dictate. The version of free will in which an agent must have more than one possible way of acting is ruled out by determinism, and the independent condition for ascribing responsibility can thus not be met unequivocally.

[18] John Martin Fischer uses the idea of 'knowingly acting' as the key to his conception of responsibility. [Fischer, 1994]

[19] John Searle notoriously claims that humans and other 'higher' sentient beings such as his dog, Ludwig Wittgenstein, have intentional minds where other entities are mere machines [Searle, 1983, Searle, 1994]. This idea has not been left unchallenged. Dennett for instance claims that intentionality is just a meaningless notion [Dennett, 1993]

[20] The concept of free will used in the ICT is the concept in which an agent has alternate possibilities. In the literature on free will several other concepts are used, e.g. that one has free will if the contents of one's will are solely determined by the agent himself, i.e. if he is autonomous in the proper sense of the word. Although Braham and Van Hees use 'autonomous agents' they intend to talk about free will as 'having alternatives'. Cf. [Frankfurt, 1971]

These arguments have deep metaphysical implications for the whole concept of moral responsibility, not just for IS. We do not know how to ascribe intentionality to IS, but neither do we know how to ascribe it to human agents. We do not know how to ascribe free will to IS, but neither do we know how to ascribe it to human agents.

Concerning human agents, the debate turns on the question whether the relation between input and output is deterministic. This debate has not abated yet and a concluding answer has not yet been found. The problem with IS however, is even more severe. Even if their input is not deterministic all present-day IS have a deterministic input-output relation (even though they can learn, their protocols are determined). This means that even if all-out causal determinism turns out to be false IS have no free will.

Intentionality is a controversial concept that we, perhaps, can circumvent by some redefinition of the independent condition in terms of susceptibility to reason, but causal determinism is firmly argued for and the implications of determinism for free will are devastating. Why are we not devastated by this thesis? Why are we still holding people morally responsible if there is a substantiated theory that means that, no matter what our intentions are, we will act the way we will act?

In 1962 Peter Strawson's essay *Freedom and Resentment* tried to answer these questions and posited a completely different concept of what it means to be a moral agent [Strawson, 1962]. He argues that moral responsibility is not to be understood as an external theoretical concept like ICT that can be applied to human interaction but is a fact founded in the practice of human social life. Therefore the threats to the theoretical concepts of intentionality and free will become irrelevant. In the next section his take on responsibility will be outlined, in section 3.2 I will investigate whether we can apply Strawson's conception of responsibility to IS.

3 Strawson: moral agents as participants in a moral community

3.1 Outline of Strawson's concept of responsibility

In *Freedom and Resentment* Strawson is not concerned with intentionality but with determinism and the question if determinism threatens free will and thus responsibility. His answer is that determinism does not threaten any conception of free will that is significant for other people than philosophers. He aims to show that assigning responsibility to agents is a practice internal to human social life and stands in no need for any external justification. His arguments will be sketched to fully understand this claim.

Strawson starts out by making a distinction between what he calls 'optimists' and 'pessimists', the former of which hold that the truth of determinism means that the concepts and practices of moral responsibility do not lose their applicability, the latter of which hold that the truth of determinism means these concepts have no application, and the practices of punishing and blaming, of expressing moral condemnation and approval, are really unjustified. The optimist can be understood as a compatibilist, the pessimist as an incompatibilist. Compatibilists hold that even if causal determinism is true then there is still a meaningful way of using free will and responsibility, the incompatibilist holds that determinism rules out free will and thus rules out responsibility [McKenna, 2009, McKenna and Russell, 2012]. The pessimist can either be a libertarian, denying the truth of determinism, or a moral sceptic, denying the validity of any moral judgement.

How can the optimist claim that determinism and a meaningful application of responsibility are compatible? He does this by using freedom in a strictly negative way[21]. This means that one is free when one is not coerced by force or innate incapacity. Strawson states:

> [T]he general reason why moral condemnation or punishment are inappropriate when these [deterministic] factors or conditions are present is held to be that the practices in question will be generally efficacious means of regulating behaviour in desirable ways only in cases where these factors are not present.
>
> *[Strawson, 2008, p. 3]*

[21] The distinction between negative and positive freedom is generally attributed to [Berlin, 1959], but is in fact much older and can be traced back to at least Kant [Carter, 2012].

Freedom is present when practices of moral condemnation and punishment are efficacious. The pessimist has to admit that this notion of freedom is compatible with determinism, but insists that

> this is not a sufficient basis, it is not even the right sort of basis, for these practices [like condemnation and punishment] as we understand them. [...] the admissibility of these practices, as we understand them, demands another kind of freedom, the kind that in turn demands the falsity of the thesis of determinism.
>
> *[Strawson, 2008, p. 4]*

The pessimist claims that the optimist is leaving out "something vital" [Strawson, 2008, p. 24], some sense of freedom that encompasses the notion of determinism. The optimist scoffs the "obscure and panicky metaphysics of libertarianism" [Strawson, 2008, p. 27] and holds that moral scepticism is untenable.

Strawson decides to introduce a completely new methodology to solve this stand-off between the pessimists and optimists. A new methodology that does not start with theoretical considerations, but with the actual practice of attributing responsibility. The approaches of the pessimists and optimists "permit, where they do not imply, a certain detachment from the actions or agents which are their objects" [Strawson, 2008, p. 5]. Instead, Strawson wants to begin with a description of human interaction:

> I want to speak, at least at first, of something else: of the non-detached attitudes and reactions of people directly involved in transactions with each other; of the attitudes and reactions of offended parties and beneficiaries; of such things as gratitude, resentment, forgiveness, love, and hurt feelings.
>
> *[Strawson, 2008, p. 5]*

By analysing resentment Strawson describes different reactions (what he calls personal reactive attitudes) we display when we are hurt by the actions of an agent. We blame the agent and feel resentment or we don't blame the agent and there is no resentment. When we blame the agent for his action we see this agent as a normal participant in our human interactions and ascribe him intentional choice in the act. When we do not blame the agent there are two possibilities: 1) we see the agent as a normal participant that is an appropriate subject of praise or blame, or 2) we exclude the agent from our moral community of appropriate subjects for praise or blame. In the first case the agent caused hurt accidentally, unknowingly, unintentionally or is

excusable by some such relevant considerations. In the second case we do not blame the agent because we perceive resentment as uncalled for because the agent is an inappropriate subject for such reactive attitudes. Such an agent is somehow unable to make informed judgements about the consequences of his actions and unable to 'intend' in a malevolent way. Young children and psychiatric patients may be such agents. In the case of agents who are not appropriate targets for praise or blame we discard the direct reactive attitude and take up what Strawson calls an "objective attitude" [Strawson, 2008, p. 10].

These two attitudes are just facts of life, Strawson claims. It is part of our nature to display these attitudes, part of our social genome. He writes: "the existence of the general framework of attitudes itself is something we are given with the fact of human society" [Strawson, 2008, p. 25]. The fact that we are social animals make us attribute (or withhold, in the case where we take the objective attitude) blame or praise and thus responsibility long before any theoretical considerations about causal determinism come into play. He has three arguments for this claim.

First: the truth of determinism cannot structurally damage our approach to agents. For if determinism is true, it determines and thus excuses all acts and all agents. If we must excuse all acts and all agents we must accept that either all agents act unintentionally or all agents are morally incapacitated. Clearly we have acts that are intentionally malevolent and clearly we have agents that are morally capable, therefore the truth of determinism doesn't structurally undermine the applicability of reactive attitudes.

Second: Strawson argues that, even if we can, on occasion, adopt an objective (detached) attitude it would be psychologically impossible for us to adopt it all of the time. Strawson:

> I am strongly inclined to think that it is, for us as we are, practically inconceivable. The human commitment to participation in ordinary inter-personal relationships is, I think, too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction might so change our world that, in it, there were no longer any such things as inter-personal relationships as we normally understand them; and being involved in inter-personal relationships as we normally understand them precisely is being exposed to the range of reactive attitudes and feelings that is in question.
>
> *[Strawson, 2008, p. 12]*

Third: it might be argued that ignoring, on a practical level, the truth

of determinism is not rational. Strawson argues that it is exactly the sense of 'rational' that is at stake here:

> It is a question about what it would be rational to do if determinism were true, a question about the rational justification of ordinary inter-personal attitudes in general. [...] And I shall reply [...] [that] we could choose rationally only in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment; and the truth or falsity of a general thesis of determinism would not bear on the rationality of this choice[p. 14]. [...] The rationality of making or refusing it would be determined by quite other considerations than the truth or falsity of the general theoretical doctrine in question. The latter would be simply irrelevant[p. 20].
>
> *[Strawson, 2008, pp. 14–20]*

What is considered rational is not a matter of the truth or falsity of some proposition, but a matter of the "gains and losses to human life". It need not be argued that rejecting any notion of responsibility would be a great loss to human life and that holding on to such a notion would be a great gain. Personal reactive attitudes are thus justifiably used in human society. Strawson goes on to argue that moral reactive attitudes, which are "generalized or vicarious analogues", rest on exactly the same arguments:

> The personal reactive attitudes rest on, and reflect, an expectation of, and demand for, the manifestation of a certain degree of goodwill or regard on the part of other human beings towards ourselves; or at least on the expectation of, and demand for, an absence of the manifestation of active ill will or indifferent disregard. (What will, in particular cases, count as manifestations of good or ill will or disregard will vary in accordance with the particular relationship in which we stand to another human being.) The generalized or vicarious analogues of the personal reactive attitudes rest on, and reflect, exactly the same expectation or demand in a generalized form; they rest on, or reflect, that is, the demand for the manifestation of a reasonable degree of goodwill or regard, on the part of others, not simply towards oneself, but towards all those on whose behalf moral indignation may be felt, i.e., as we now think, towards all men.
>
> *[Strawson, 2008, pp. 15–16]*

Just as "the very great importance that we attach to the attitudes and intentions towards us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions" [Strawson, 2008, p. 5] compels us to accept the validity of our personal reactive attitudes, so the commitment to moral reactive attitudes validates those attitudes. It is inconceivable, argues Strawson, that we abandon our ascriptions of blame or praise, of moral responsibility, in the face of mere theoretical considerations.

Both the optimist and the pessimist miss an important point. Both seek arguments for attributing responsibility based on objective attitudes, both disengage from the direct practice of normal human interaction.

> Both seek, in different ways, to over-intellectualize the facts. Inside the general structure or web of human attitudes and feelings of which I have been speaking, there is endless room for modification, redirection, criticism, and justification. But questions of justification are internal to the structure or relate to modifications internal to it. The existence of the general framework of attitudes itself is something we are given with the fact of human society. As a whole, it neither calls for, nor permits, an external 'rational' justification. Pessimist and optimist alike show themselves, in different ways, unable to accept this.
>
> *[Strawson, 2008, p. 25]*

What Strawson demonstrates is that moral responsibility is determined not by external, that is, detached, considerations, arguments and theories, but by practices that are internal to any moral community. These practices are simply a given and natural part of participation in human society.

In *Freedom and Resentment* Strawson dismisses the sceptical conclusion that since we do not know whether there is free will or not, and we do not know whether anyone ever acts intentionally or not, we cannot ascribe responsibility to anyone. The "inescapable psychological mechanisms that guide human thought and action" [McKenna and Russell, 2012] give reason to meaningfully apply the concepts of blame, praise and responsibility. To fully appreciate this point, imagine waking up one day and reading that causal determinism is proven beyond a shimmer of a doubt to be true. What would this mean for the practice of our lives and our society? The consequences would be negligible for we would still live our lives as meaningful projects and treat others as being part of those projects and having meaningful projects themselves. As Pereboom states: "Although hard

incompatibilism would diminish the sense in which we can have genuine achievements, and the sense in which we can be worthy as a result, it would by no means thoroughly undermine the fulfillment in life that our projects can provide." [Pereboom, 2003, p. 188]

Freedom and Resentment has been criticized. Critics have suggested that the main argument rests on a version of the naturalistic fallacy. The mere fact that we would attribute responsibility in the face of the truth of determinism is not in itself a validation of that fact. Historically, we have retracted many convictions we once held dear on the basis of new arguments and information, so why wouldn't we be mistaken on the subject of the appropriateness of assigning responsibility? Strawson would reply that one is at liberty to accept the consequences of determinism but then one must also accept the fact that no-one is ever to blame and that reward, punishment etc are vacuous. Surely, Strawson would suggest that this goes against the very fabric of human interaction and thus against the fabric of human life.

Moreover, the criticism aimed at *Freedom and Resentment* is not of special relevance here since that criticism is aimed at the validity of attributing responsibility to human agents and not at the applicability of these arguments to IS. In the remainder of this essay Strawson's take on responsibility is used as it stands, keeping in mind that criticism aimed at *Freedom and Resentment* will in all probability also apply to the analysis in the following section.

3.2 Strawson's conception of moral responsibility applied to IS

In section 2.1 of this essay the independent condition theories, ICT, were introduced and tested. These theories try and explicate conditions that must be met if we are to assign responsibility (assigning blame or praise) to an agent. It turned out that it is impossible to apply ICT to IS, because we do not know whether the conditions are satisfied. The AOC Braham and Van Hees use, for instance, requires the existence of free will as the opportunity to do otherwise, and the possible truth of the thesis of determinism makes it unclear whether such free will exists. The AC is equally unclear because we do not know what acting intentionally means or how we could verify whether an agent acted intentional or not.

In section 2.4 it was discussed that those problems hold for more than just questions surrounding responsibility and IS because those problems hold for human agents as well, although the question about the determinism of

the input-output relation was easier to settle for present-day IS. *Freedom and Resentment* was introduced as a source in which the problems that the possible truth of determinism holds for human responsibility are confronted by a bottom-up approach. The fact that we do, and rationally so, attribute responsibility to human agents renders purely theoretical claims about determinism moot. If the fact that we cannot know whether human agency satisfies AOC or AC is made irrelevant for the attribution of responsibility to humans, then it might be the case that the fact that we cannot know whether IS agency satisfies AOC and AC is also irrelevant for the attribution of responsibility. If human responsibility needs no external (independent) conditions, why should IS responsibility need it?

As described, Strawson uses the psychological facts of human nature to argue for the validity of assigning blame or praise to agents. In normal circumstances a human agent is seen as a member of a moral community in which his actions evoke reactive attitudes. In some cases the initial reaction is revoked and replaced by an objective attitude because it is conceded that the agent did not act intentionally and is thus not worthy of praise or blame for that specific act. In other cases the agent is seen as external to the moral community and not a suitable target for praise or blame because he is not able to do anything intentionally (because he is unable to judge, reason, weigh consequences and so forth). Such an agent should therefore be approached solely with an objective attitude, as something which can be understood and controlled, but which does not have proper intentions.

The main question of this section is whether an IS is, in Strawson's analysis, a proper target for blame or praise, or one of those agents which are not really agents because they do not know what they are doing and only satisfy condition Braham and Van Hees' CRC. The question we should ask ourselves is "Can we understand an IS as a member of our moral community?".

Interestingly, the question is not whether an IS is intentional, conscious, autonomous, etc., but whether an IS can, would or should be treated as if he were one of us, a participant of our moral community. This is a markedly different question than the questions that form the main discussions on AI and responsibility. They almost invariably turn on these intractable properties or entities human acts are supposed to possess. Searching for these properties in order to be able to know when to attribute responsibility to IS is searching for external conditions all over again[22]. Strawson's take

[22]Although several AI-projects focussing on embodied and embedded cognition are using a kind of socialisation-theory in which an IS is developing like a child develops. Researchers are expecting that this will lead to IS that are very similar to normal humans and could thus be included in our community. For instance the COG-project at

bypasses these questions and aims at the practical side. Would we identify an IS as a member of our moral community? This question can be seen as the question whether an IS would pass a generalized Turing test[23], not aimed at conversational intelligence, but aimed at whatever it takes to be accepted as a member of our moral community.

MIT, http://www.ai.mit.edu/projects/humanoid-robotics-group/cog/cog.html and http://users.ecs.soton.ac.uk/harnad/Papers/Py104/dennett.rob.html

[23]Thanks to dr. Dyrkolbotn for the remark on the generalised Turing Test

4 Moral membership of IS

4.1 Six preliminary observations for future research

The question whether we would identify an IS as a member of our moral community is ultimately a sociological/psychological[24], one like how we discern morally capable agents from children and the insane. Strawson admits in *Freedom and Resentment* that his dichotomy of reactive and objective attitudes is rough and many intermediate situations could and should be acknowledged. For example, there is no exact moment at which a child turns from innocent angel to fully-fledged moral agent. A psychological analysis is beyond the scope of this essay but I will make a number of preliminary observations that might guide such an analysis in this section.

Observation 1 Humans have the ability to discern morally capable agents from morally incapable agents. This ability is not flawless or all-powerful because we sometimes make mistakes when attributing or withholding responsibility. E.g. when we wonder whether to assign moral capability to criminals. We also encounter situations in which we cannot really tell the difference, e.g. with children.

Observation 2 Moral capability or blameworthiness and praiseworthiness is not an all-or-nothing affair. Human agents differ in their (cognitive, social, etc.) abilities and whether they are blamed or not depends for a great deal on those abilities. For example, if I (not a physician) make a horrible medical mistake while trying to save a dying man I will not be blamed for I do not have the know-how needed to discern between the right treatment and the mistake. In this case, as in others, I would even be obliged not to help. A trained doctor will be blamed for doing exactly the same thing I did, because he has the know-how and isn't expected to make such a horrible medical mistake.

Observation 3 If an agent is blamed or not depends partially on the judgement of the capacities of that agent. It is the (informed) judgement of the community which determines whether the agent is to be blamed or not. E.g. it is said that the agent 'should have known better' even if he did not know better.

Observation 4 Humans have a remarkable tendency to assign responsibility to anything that does not in the slightest resemble a human. Most

[24] A psychological or sociological question because it is aimed at understanding how individuals or groups accept or reject entities like agents or concepts.

people have blamed their computer, their car, the weather, etc. for bringing about all kinds of misery. Apparently sometimes (although most of the time only for a moment) we believe these machines and phenomena cause the misery, have the opportunity to do something else and act intentionally (perhaps even cruelly?).

Observation 5 The more complex or inscrutable the agent we encounter is, the easier it is for us to assign to the agent all sorts of intentions. When my ballpoint fails it's not easy to see how I could blame it, but when something as complex as my notebook fails I sometimes blame it for failing. When a severely mentally handicapped human injures me I'm hard pressed to feel resentment, but when an intelligent person hurts me I will be angry at him.

Observation 6 We excuse agents that we do not consider to be full members of our moral community. Depending on their level of development in our moral community we assign or withhold judgement, as in the case of a stranger that is not accustomed to our ways, as in the case of a child that does not yet know the subtle rules we live by.

What requirements of inclusion in our moral community emerge from these observations? It must be noted that we are once again looking at conditions, this time the psychological conditions for judging an agent as an individual on a par with his judges. Taking the six observations the following six statements can be made:

First statement Physical resemblance is not a requirement.

Second statement The community decides which agent in which situation is worthy of blame or praise, thus tailoring moral capacity to dominant culture and tailoring it to what abilities an agent is expected to exhibit.

Third statement Our judgement of moral capacity is flawed and dynamic, i.e. we make mistakes and alter our judgements according to new information and shifts in theoretical frames. There even are agent-acts[25] which we cannot decide to categorize as a proper target of moral scrutiny or not. Should a 5-year old know better than to try and wash the car with sand? We do not blame him, but do reprimand him in order to raise him to full moral membership.

[25] I introduce the concept of agent-act to emphasize that we can morally judge an act only when we take the agent who acted into consideration.

Fourth statement We are not tied to the view that humans and only humans can be moral agents. Of course we differentiate between full moral agency and less-than-full moral agency, and hesitate to attribute full membership to non-humans, but the wall that separated humans and non-humans in terms of rights and responsibilities has been breached.[26]

Fifth statement Complexity or inscrutability is a requirement.

Sixth statement We tend to assign full responsibility only to those agents we believe to be full members of our moral community. Less-than-full membership implies less-than-full responsibility. The second observation determines the expected developmental membership.

4.2 Statements on the possible moral membership of IS

What do these observations tell us about the relation between IS and our moral community? I will try and answer this question by using three statements which I consider false, based on the six preliminary observations and six statements formulated in section 4.1.

- IS are too different from humans to be included in the human moral community.

Observation 4 claims that we sometimes assign responsibility to machines and phenomena that are distinctly non-humanoid. Although getting angry at your desktop is quickly dismissed as ludicrous, praising the dog is considered normal and even functional behaviour. After getting angry at your pc, you realize that the pc does not have any kind of judgement, moral or not, and you dismiss your anger. In the case of dogs (or other creatures that we think of as capable of learning through blame or praise) we insist in our judgement. Apparently the dog should or could know better, is susceptible to reason (even if it is reasoning on the level of direct fear/pain or joy/pleasure) and is thus treated and thus recognized as a member of our moral community. He will never be a full member and will not be considered to be responsible for making complicated decisions, but he will be trained, punished and rewarded.

An IS, even if non-humanoid, will have to be considered a member of our moral community, whose membership is dependent on the measure in which

[26] For instance, the Volkskrant reported on December 23rd on rights granted by a judge to an orang-utan named Sandra. `http://www.volkskrant.nl/buitenland/erkenning-grondrechten-mensaap-baant-de-weg-voor-andere-dieren~a3816523/`

he is judged to be able to reason. That is, if he is able to learn through feedback and other responses, weighing reasons directed at goals (even it is on the level of direct fear/pain or joy/pleasure), and performing acts based on that reasoning. This means IS could fit statements 1 and 4 from the previous section.

- IS are completely determined by their code and can thus never be regarded as moral agents.

As long as IS are complex or inscrutable enough one of the requirements is met: it is not obvious how the act or failure to act is caused by constitutive preceding situations, mechanical causality or unavoidable laws of nature, therefore suggesting a measure of free will[27]. In such cases we are wont

[27] John Martin Fischer, following Harry Frankfurt's suggestion that free will consists in doing what you really want to do [Frankfurt, 1971, p. 18], claims that guidance control is sufficient for ascribing free will. Even if there is no total control or regulatory control (needed for the ability to want otherwise or to do otherwise, respectively), ascribing guidance control is all that is needed to assign responsibility. Fischer calls this "reasons-responsiveness" [Fischer, 1998, p. 222]. Dennett lays down a similar claim in [Dennett, 1984a] and [Dennett, 1984b]. Dennett:

> "If it is unlikely that it matters whether a person could have done otherwise - when we look microscopically closely at the causation involved - what is the other question that we are (and should be) interested in when we ask "But could he have done otherwise?"? Consider a similar question that might arise about a robot, destined (by hypothesis) to live its entire life as a deterministic machine on a deterministic planet. Even though this robot is, by hypothesis, completely deterministic, it can be controlled by "heuristic" programs that invoke "random" selection - of strategies, policies, weights, or whatever - at various points. All it needs is a pseudo-random number generator, either a preselected list or table of pseudo-random numbers to consult deterministically when the occasion demands or an algorithm that generates a pseudo-random sequence of digits. Either way it can have a sort of bingo-parlor machine for providing it with a patternless and arbitrary series of digits on which to pivot some of its activities. Whatever this robot does, it could not have done otherwise, if we mean that in the strict and metaphysical sense of those words that philosophers have concentrated on. Suppose then that one fine Martian day it makes a regrettable mistake: it concocts and executes some scheme that destroys something valuable-another robot, perhaps. [...] It does not matter for the robot, someone may retort, because a robot could not deserve punishment or blame for its moments of malfeasance. For us it matters because we are candidates for blame and punishment, not mere redesign. You can't blame someone for something he did, if he could not have done otherwise. This, however, is just a reassertion of the CDO [Could Do Otherwise] principle, not a new consideration, and I am denying that principle from the outset. Why indeed

to ascribe freedom of the will to an agent. As long as his actions are not overtly caused by 'external' factors we assign the agent at least some form of autonomous agency in the act. If we are not checked by theoretical considerations whether an agent is really praiseworthy or blameworthy our attitudes are reactive in Strawson's sense of the term. The IS as an agent meets this requirement of membership. Remember that in the first section it is argued that intelligent systems are unpredictable and their acts are not necessarily determined more so than human acts are. This means IS could fit statement 5 from the previous section.

- Humans will never allow IS as members of the human moral community.

The check that turns our attitude form reactive to objective depends on ideas that a community has on moral agency. Evaluating what agent-act is a suitable target is a part of the way we organize our life world, and determines our practices and is determined by them, informed but not governed by theoretical considerations. Statement 2 points to this aspect of moral judgement by emphasizing the role of the community in deciding which agent-act is praiseworthy or blameworthy. However, as a moral community we change the way we judge an agent's blameworthiness and praiseworthiness, based on developments in theoretical standards, but also based on all the factors influencing all of our social practices. Which agent-act we deem worthy of our praise or blame is thus not static.

Where statement 2 states the dynamics of how responsibility is assigned in general, statements 3 and 6 focus on particular judgements. They both point toward the ways we differentiate between agents based on their cognitive and moral development. Apparently it is not always obvious when to assign responsibility and when to withhold such judgement. A certain amount of fallibilism is part of our moral culture. Furthermore, we accept that agents tend to develop or have the potential for moral growth. Not one of us is born as a fully-fledged moral agent, and not one of us can seamlessly adjust to the moral expectations of an entirely different culture. Humans need time to adapt and to learn, as kids growing up or as adults adjusting when expatriated. This is accepted as fact in our moral community, and

---

shouldn't you blame someone for doing something he could not have refrained from doing? After all, if he did it, what difference does it make that he was determined to do it?'
[Dennett, 1984b, pp. 559-563]

part of the way we assign responsibility. Eventually all normal adults are expected to act as full members of our community (excepting those adults that are deemed unable to reason and are thus deemed unfit targets for moral appraisal[28]), but everybody gets allowed a training period.

The allowance of a training period is perfectly acceptable for humans, and we are quite able to differentiate between full members and members-in-training. We might also be speciesist (to borrow a phrase made famous by Peter Singer[29]), but we also tend to stretch the line between humans and non-humans (see statement 4). When we combine these observations or statements we can come to the following statement; if IS get the chance to learn, to develop as a member-in-training, they can acquire the same status human learners get granted. It might even be argued that the main difference between cognitively and operationally equivalent IS and humans is the degree to which they are immersed in and formed by a moral community. Without training or guidance IS might not develop any moral motivations or ethical skills and remain more distant from our moral community than a space alien. As [Bostrom, 2014] writes, the descent from a social (and moral) community determines whether an agent can ever acquire social and moral skills:

> An artificial intelligence can be far less human-like in its motivations than a green scaly space alien. The extraterrestrial (let us assume) is a biological creature that has arisen through an evolutionary process and can therefore be expected to have the kinds of motivation typical of evolved creatures. It would not be hugely surprising, for example, to find that some random intelligent alien would have motives related to one or more items like food, air, temperature, energy expenditure, occurrence or threat of bodily injury, disease, predation, sex, or progeny. A member of an intelligent social species might also have motivations related to cooperation and competition: like us, it might show in-group loyalty, resentment of free riders, perhaps even a vain

[28] It is probably a major part of being normal that one is fit for moral judgement. As the community decides what is morally right and wrong the normality of its members is determined by their fit in this moral spectrum. Michel Foucault analysed these mechanisms through which people get disciplined and judged to normality in several of his books, most notable in History of Madness. [Foucault, 1961]

[29] Peter Singer did not come up with 'speciesism', the equivalent of racism based not on race but on species, but is the most well-known user of the term, e.g. in *Animal Liberation: A New Ethics for our Treatment of Animals* [Singer, 1977].

concern with reputation and appearance.

> An AI, by contrast, need not care intrinsically about any of those things. There is nothing paradoxical about an AI whose sole final goal is to count the grains of sand on Boracay, or to calculate the decimal expansion of pi, or to maximize the total number of paperclips that will exist in its future light cone. In fact, it would be easier to create an AI with simple goals like these than to build one that had a human-like set of values and dispositions. Compare how easy it is to write a program that measures how many digits of pi have been calculated and stored in memory with how difficult it would be to create a program that reliably measures the degree of realization of some more meaningful goal human flourishing, say, or global justice. Unfortunately, because a meaningless reductionistic goal is easier for humans to code and easier for an AI to learn, it is just the kind of goal that a programmer would choose to install in his seed AI if his focus is on taking the quickest path to "getting the AI to work" (without caring much about what exactly the AI will do, aside from displaying impressively intelligent behavior).
>
> *[Bostrom, 2014, p. 106–107]*

As mentioned in section 4.1 the question whether a moral community can accept IS as members, let alone full members, is ultimately a psychological question. It remains to be seen if IS, given proper formative years and guidance as members-in-training, will get accepted as part of our moral community.

It seems that statements 2, 3 and 6 depend on psychological considerations that are beyond the scope of this essay. It can be observed, however, that the psychology of our moral community has shown itself to be dynamic, accepting new members and rejecting old ones, based on theoretical insights, new information and slowly changing views on life and live forms, seeping through to our everyday practice of assigning responsibility. It is likely that IS that are raised as members of our community one day will be granted a full membership, notwithstanding all kinds of theoretical worries about underlying differences with human members.

Conclusion

When trying to figure out which agent-act is a suitable target for assigning responsibility ICT, independent condition theories, are brought to the fore. Applying these to human agents demands applying the concepts of intentionality and autonomy, concepts which are thought of as not applicable to IS. Therefore IS are thought of as unsuitable targets for assigning responsibility. The concepts used on human agent-acts are troublesome, however. From a philosophical perspective they are – at least at the moment – impossible to define and thus impossible to predicate. The question whether we can understand responsibility at all, let alone in IS, comes up. ICT makes demands we cannot meet, so where can we turn? Peter Strawson tries out a new turn[30] in thinking about responsibility by placing the determining force in the community. It is the practice of everyday life that determines responsibility. This take bypasses the philosophical problems encountered in applying ICT and uses the concept of moral community as a criterion. The question then becomes whether IS can become members of our moral community so we can assign them responsibility. It turned out that, although several preliminary observations could be made and be reformulated as statements, the answer to this question can only be given by psychological research. Such research could be one of the main challenge in robotics. How can we adjust to IS? IS are entering our everyday life, and are not simple machines we can put next to our hammer, lawnmower and dishwasher. As IS come ever closer to cognitive and functional equivalency with humans we will have to reorganise the admissions committee of our moral community.

[30] Although it could be argued that his conception of responsibility harks back to ancient Greek, e.g. Aristotle's, thoughts about how responsibility is a concept applicable only to man as *zoion politikon*.

References

[Asimov, 2004] Asimov, I. (2004). *I, robot*. Random House LLC.

[Berlin, 1959] Berlin, I. (1959). Two concepts of liberty: An inaugural lecture delivered before the university of oxford on 31 october 1958.

[Bostrom, 2014] Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

[Braham and Van Hees, 2012] Braham, M. and Van Hees, M. (2012). An anatomy of moral responsibility. *Mind*, 121(483):601–634.

[Broersen, 2014] Broersen, J. (2014). Responsible intelligent systems. *KI - Knstliche Intelligenz*, 28(3):209–214.

[Carter, 2012] Carter, I. (2012). Positive and negative liberty. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2012 edition.

[Chalmers, 1995] Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200–219.

[Dennett, 1984a] Dennett, D. C. (1984a). *Elbow room: The varieties of free will worth wanting*. MIT Press.

[Dennett, 1984b] Dennett, D. C. (1984b). I could not have done otherwise–so what? *The Journal of Philosophy*, 81(10):553–565.

[Dennett, 1993] Dennett, D. C. (1993). *Consciousness explained*. Penguin UK.

[Dennett, 1997] Dennett, D. C. (1997). When hal kills, who's to blame? computer ethics. In Stork, D.G., e. a., editor, *HAL's Legacy: 2001's Computer as Dream and Reality*. Cambridge, MA: MIT Press.

[Fischer, 1994] Fischer, J. M. (1994). The metaphysics of free will: A study of control.

[Fischer, 1998] Fischer, J. M. (1998). Moral responsibility and the metaphysics of free will. *The Philosophical Quarterly*, 48(191):215–220.

[Foucault, 1961] Foucault, M. (1961). History of madness.

[Frankfurt, 1971] Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68(1):5–20.

[Matthias, 2004] Matthias, A. (2004). The responsibility gap - Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6:175–183.

[McKenna, 2009] McKenna, M. (2009). Compatibilism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2009 edition.

[McKenna and Russell, 2012] McKenna, P. and Russell, P. (2012). *Free Will and Reactive Attitudes: Perspectives on P.F. Strawson's "Freedom and Resentment"*. Ashgate Publishing, Limited.

[Moor, 2006] Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *Intelligent Systems, IEEE*, 21(4):18–21.

[Nagel, 1991] Nagel, T. (1991). *Mortal Questions: Canto*. Cambridge University Press.

[Noorman, 2014] Noorman, M. (2014). Computing and moral responsibility. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2014 edition.

[Pereboom, 2003] Pereboom, D. (2003). *Living without free will*. Cambridge University Press.

[Searle, 1980] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(03):417–424.

[Searle, 1983] Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.

[Searle, 1994] Searle, J. R. (1994). Animal minds. *Midwest studies in philosophy*, 19(1):206–219.

[Singer, 1977] Singer, P. (1977). *Animal liberation. Towards an end to man's inhumanity to animals.* Granada Publishing Ltd.

[Strawson, 1962] Strawson, P. F. (1962). Freedom and resentment.

[Strawson, 2008] Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.