Strawson's take on responsibility applied to AI

Laura Cromzigt BSc
Utrecht University
lauracromzigt@gmail.com

Summary of `Strawson's take on moral responsibility applied to Intelligent Systems' for the Student Research Conference 2016

ABSTRACT
This paper investigates the attribution of responsibility to artificial intelligent systems. It argues that traditional approaches to the subject are insufficient because they encounter some of the same problems that one encounters when attributing responsibility to humans. Peter Strawson's take on responsibility is introduced as an alternative approach. He claims that theoretical considerations miss the point when we ponder the responsibility of human agents. Instead, we should understand responsibility as part of the practice of human life. This claim is investigated and transferred to AI to see if it provides a more fruitful way to understand responsibility of artificial intelligent systems.

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

INTRODUCTION

As artificial intelligent systems (AI, in this paper also used for the singular) become more and more advanced, questions about responsibility arise. For example, who is responsible when AI cause an accident? AI, such as drones and driverless cars, are considered learning (i.e. semi-autonomous) automata and are thus not merely tools in the hands of human agents, fully controlled by the operator [Matthias, 2004]. As AI become part of our human world more and more 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. This paper will look into what the traditional way of thinking about responsibility and AI amounts to, why it does not succeed in ascribing responsibility to AI and whether Peter Strawson's approach to humans as members of a moral community is worth further investigation in relation to AI.

INDEPENDENT CONDITION THEORIES

The most common way responsibility is attributed to humans, which I will call independent conditions theories (IC theories), will be investigated first.

Conditions

IC theories expect that you have to fulfill certain conditions in order to be responsible. These three conditions are 1) autonomy, 2) having reasonable alternatives and 3) being causally relevant. 1) Autonomy is commonly held to be the condition of being able to intentionally develop reasons and act upon those reasons. It sometimes gets discussed when we are concerned with developing children and the sanity of grown-ups. 2) Whether someone had reasonable alternatives for his actions depends on the definition of reasonableness. This definition sometimes gets discussed in real life when trying to discern between temptation and compulsion. 3) Whether one's action is causally relevant to some event becomes interesting when discussing global causal connections, such as one's contribution to global warming. An example an IC theory will be sketched first before turning to the applicability to AI.

Example IC theory

Braham and Van Hees (2012) think that the three conditions are independently necessary conditions, meaning that each is a necessary but not sufficient reason to assign moral responsibility:

- 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.

It will be argued that it cannot be ascertained whether such conditions apply to AI.

IC theories applied to AI

When applying IC theories to AI it needs to be checked if we can verify all three conditions. In this summarized paper I will only describe the Agency Condition (AC), although similar arguments can be developed against the Avoidance Opportunity Condition (AOC).

Agency Condition

An AI or human satisfies this condition if it performed the act autonomously. This means that the AI or human has to be aware 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. Moreover, it has to be the uncaused source of its own considerations. How can we tell in the case of an AI? Does an AI which acts as if it knows what it is doing and which gives reasons for what it is doing really know and really have its own reasons? In the lively debate between strong and weak AI such questions are crucial. Proponents of AI as truly intelligent systems, such as Dennett, claim that acting as if it is reasonable is reason enough to attribute autonomous reasonableness to a system, opponents, e.g. Searle and Chalmers, claim that such acting is nothing but a hollow shell, an imitation of true reasonableness lacking some vital ingredient to be

SRC 2016, The Netherlands.

rightfully called reasonable. The problem goes deeper: when trying to figure out why humans act the way they do we are confronted by the puzzling fact that we cannot tell whether humans are autonomous, because we cannot tell whether their acts are initiated by themselves or initiated by other factors. In other words, causal determinism might be true. The thesis of causal determinism claims that all of physical reality, including humans, is governed by laws of cause and effect and therefore we cannot tell for sure whether a human can be held responsible. There is as yet no clear solution to this problem. As long as this stays unanswered we do not know whether an AI (or human) meets the AC.

The IC theories way of thinking about responsibility is thus not sufficient when reasoning about responsibility and AI because 1) it cannot be ascertained whether an AI or human meets the AC and 2) all the conditions need to be met for ascribing responsibility.

Are we then left empty-handed? Do we have to develop a firm understanding of what it means to apply theoretical concepts to human actions before we can decide on the applicability of responsibility to AI? Is a 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 situations in which we intuitively know when an agent is to be held responsible? It is clear from these findings that we need to consider new approaches for thinking about assigning responsibility to AI because as long as we cannot fully understand what `acting intentionally' (AC) or for that matter what `having a reasonable opportunity to do otherwise' (AOC) mean, responsibility cannot be attributed to AI. Peter Strawson developed such a new approach.

STRAWSON ON RESPONSIBILITY

In *Freedom and Resentment* Strawson (2008) demonstrates that moral responsibility is not determined by external, that is, detached, considerations, arguments and theories like IC theories, but by practices that are an internal part of any moral community. These practices are simply a given and natural part of participation in human society. Therefore the threats of causal determinism (his main challenge in *Freedom and Resentment*) to the theoretical concepts of autonomy, intentionality and free will become irrelevant. Strawson uses the psychological facts of human nature to argue for the validity of assigning blame or praise to agents. In normal circumstances we have a so-called reactive attitude where a human agent is seen as a member of a moral community. We are simply prone to these reactive attitudes, direct unreflected reactions, towards acts we judge as blame- or praiseworthy. In some cases this initial reaction is revoked and replaced by a so-called objective attitude, in which the agent is viewed in a more detached and objective way, 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 it is not able to do anything intentionally (because it is unable to judge, reason, weigh consequences and so forth). Such an agent should therefore be approached solely with a so-called objective attitude, as something which can be understood and controlled, but which does not have proper intentions.

These two attitudes, objective and reactive, 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 or autonomy 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. 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.

STRAWSON APPLIED TO AI

Freedom and Resentment is 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 and thus autonomy 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 AI satisfy AOC and AC is also irrelevant for the attribution of responsibility to AI. If human responsibility needs no external (independent) conditions, why should AI need it?

The questions surrounding responsibility and AI in Strawson’s analysis would thus be whether an AI is a proper target for blame or praise (using the reactive attitude), or one of those agents which are to be approached with a objective attitude. Interestingly, the question is not whether an AI is intentional, conscious, autonomous, etc., but whether an AI can, would or should be treated as if it were one of us, a participant of our moral community. This is a markedly different question than the questions that form the contemporary main discussions on AI and responsibility. They almost invariably turn on such 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 AI is searching for external conditions all over again. Strawson’s take bypasses these questions and aims at the practical side. Would we identify an AI as a member of our moral community? This question can be seen as the question whether an AI would pass a generalized Turing test, not aimed at conversational intelligence, but aimed at whatever it takes to be accepted as a member of our moral community.

Membership moral community

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-edged moral agent. A thorough sociological or psychological analysis to determine when exactly a community accepts an agent as one of its own is beyond the scope of this essay but I will make six preliminary observations that might guide such an analysis:

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 in a medical way. 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 trust in the moral capabilities and thus the applicability of responsibility to anything that does not in the slightest resemble a human. Most 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 the agent all sorts of intentions. When my ballpoint fails it is 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 am hard pressed to feel resentment, but when an intelligent person hurts me I will be angry with 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.

AI as members of our moral community

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 these observations it seems membership of our moral community, and thus being an accepted target for reactive attitudes, hinges not on being human or even being humanoid, but on the ascription of being able to learn from blame or praise. As long as the actor is seen as a complex agent, with intentions, reason-receptiveness and the ability to adjust its behavior when prompted we are able to allow anyone or anything a trial membership, giving it time to develop into a full membership.

Notwithstanding the limits of this paper it can be stated that AI could be held to be responsible, just as humans are held to be responsible, if they are accepted as members of our moral community. What it takes to be granted this membership is a question for future sociological/ psychological research.

CONCLUSION

When trying to figure out which act is a suitable target for assigning responsibility IC theories, 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 AI. Therefore AI are thought of as unsuitable targets for assigning responsibility. The concepts used on human 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 AI, comes up. IC theories make demands we cannot meet, so where can we turn? Peter Strawson tries out a new turn in thinking about responsibility by placing the determining force in the moral community. It is the practice of everyday life that determines responsibility.

It turned out that, although several preliminary observations could be made, the answer to this question can only be given by psychological research. Such research could be one of the main challenges in robotics.

How can we adjust to AI? AI are entering our everyday life, and are not simple machines we can put next to our hammer, lawnmower and dishwasher. As AI come ever closer to cognitive and functional equivalency with humans we will have to reorganize the admissions committee of our moral community.

ROLE OF THE STUDENT

L. Cromzigt was an undergraduate student working under the supervision of dr. J.M. Broersen when researching and writing this thesis. The supervisor is working on the REINS project on responsible intelligent systems where the student worked as a research assistant. Drs. M.M. van Calcar suggested to think about responsibility and AI in a totally new way which led to this research proposal. The student formulated the research question, carried out the research and wrote the paper. So far no research on Strawson's conception of responsibility and AI has been published. In this thesis a first step to a new way of thinking about responsibility and AI is taken.

REFERENCES

1. Braham, M. and Van Hees, M. (2012). An anatomy of moral responsibility. *Mind*, 121(483):601-634.
2. Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200-219.
3. Dennett, D. C. (1993). *Consciousness explained*. Penguin UK.
4. 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.
5. Matthias, A. (2004). The responsibility gap - Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6:175-183.
6. Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(03):417-424.
7. Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind.* Cambridge University Press.
8. Searle, J. R. (1994). Animal minds. *Midwest studies in philosophy*, 19(1):206-219.
9. Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.