ARTIFICIAL FREE WILL:
THE RESPONSIBILITY STRATEGY AND ARTIFICIAL AGENTS

Sven Delarivière

1. Introduction

A traditional notion of free will is often pointed out to be inherently incompatible with determinism. Nonetheless, I will argue for an account of free will which is compatible with determinism. Objections have been made (e.g. Harris, 2013, 2014) that any interpretation of free will compatible with determinism would fail to capture what is attractive about it and that we should therefore simply throw the notion out our ontology. However, I will defend that a compatibilist account is “a variety of free will worth wanting” (Dennett, 2004, pp. 132). The account I will defend and the approach I take is my own, although it will be clear that Daniel Dennett’s views have greatly influenced my thinking. To lay out my account of free will, I shall first perform some groundwork by defining freedom and relating that to ethics. With that groundwork, I can explicate how my account of free will fits with the two theses of free will: (i) that the agent could have done otherwise and (ii) that she is the proper source of her actions. With this account of free will, I will retrace my steps with an eye to Artificial Intelligence, hoping to show that my account of free will is not restricted to human agents.

2. Foundational Groundwork

2.1 Defining Freedom

Before we plunge into any account of free will, I believe it is important to explicate the (or a) notion of freedom. Often, the notion of freedom is linked to the notion of control and I believe this link is a valuable one. However, we cannot *equate* freedom with control lest we run into infinite regress problems. If a pizza-addiction starts to get you in trouble with your health, you want to be able to control your interest in food. And if your interest in controlling your food is itself becoming unhealthy, you want to be able to control that too, but we can’t (and indeed don’t) require people to control their control their control and so forth ad infinitum.

Freedom is a much more fruitful concept if we define it as attaining what is of value. Counterintuitively, this means that one can be free even if one can’t do anything at all, as long as it is possible to attain or receive what is of value. This may seem strangely passive at first, but I believe the only worthwhile reason we consider an active component of control over passively getting what is of value is because control is generally more effective to that purpose. This same wisdom is captured in the famous idiom “If you give a man a fish, he will eat for a day, but if you teach him how to fish, he will eat for a lifetime.” After all, the world is not static and certainly not designed to keep us safe and happy. Therefore, it would be beneficial to one's freedom to be designed to have capacities of keeping yourself safe and happy. However, it is important to mention that we only consider these active capacities to the extent that a certain valuable state is, or could be, in danger of faltering *without* any active involvement. For instance, as long as fish are expected to be plentiful (seemingly infinite in supply), no addition to the idiom will be made about teaching a man to breed or, more drastically, create fish. Attaining what is of value, then, is at its stablest and most effective via a capacity to achieve what is of value. This stops the infinite regress we got by defining freedom in terms of control.

We can't have control ad infinitum, but we can have control ad bonum. What is undesirable about addiction is becoming insensitive to what is of value to us. It is only to the extent that our lack of sensitivity and control may make us less capable of achieving that which is of value that we consider it important to our freedom. Consequently, the requirement for freedom to be uninvolved with determinism ("free from determinism") has no place in our demands for a variety of free will worth wanting unless it stops us from attaining what is of value. In fact, without determinism, it is hard to see how one can have a reliable capacity to (re)act appropriately in any situation.

2.2 Ethics

Since we have defined freedom as the ability to achieve what is of value, and considering morality is a mode of conduct concerned with that achievement, it should be of no surprise that morality and free will are deeply connected. If our choices bear no consequences of any actual (or perceived) value (to you or anyone else), it would be hard to call them free in any appropriate sense. We rightfully call a vote in a rigged election an illusion of choice despite our ability to vote with no one forcing us to choose one party over the other. If it doesn't effectuate any valuable (actual/perceived/possible) difference, then any difference it does effectuate we would only call freedom because of an abstract but entirely useless conception of the term.

It is my contention that, for free will (and morality in general), all significant moral arguments are either directly or indirectly *utilitarian* in nature. What this means is that right and wrong are considerations of consequences, and that morality consists of the efficient (though complex) moral strategies we employ to shape consequences into those that are of value. (Singer, 2011; Franklin, 1968) What I called "moral strategies" are the (attempted) ways to increase what is of value in a community of subjects. Which strategy is to be employed depends on its cost-effectiveness, where cost stands for both the harm it does and the difficulty of deploying it. The efficacy of each strategy varies with the dispositions of its target. Some such strategies are: destruction, isolation, punishment or reward. To the extent that an agent's intentions determine her actions, she can be the proper target of the most cost-effective moral strategy: the *responsibility strategy* (coined here). Its use is justified by the efficacy of people's ability to respond to moral reactive attitudes (and their anticipation) and/or realizing the rationale for the strategy. In short, this means being able to respond positively to being held responsible, be it for emotional reasons or by seeing the moral utility in being held responsible. Consequently, simply sharing the rationale for acting (or refraining from acting) would, when possible, be the most cost-effective strategy.

3. Could Have Done Otherwise

3.1 Analysis of Can and Otherwise

With the groundwork behind us, I believe we can tackle the two thesis of free will and their relation to determinism more adequately. The first thesis of free will requires the ability to *could have done otherwise* (henceforth CHDO). This does not seem to be allowed by determinism since a determined agent can only do what it is determined to do. However, the argument stands or falls with what is required of "can" in "could have done otherwise".

Let me first point out that notions like "abilities", "otherwise" and "similarity" are all of a higher ontological category. We are not defining abilities (e.g. playing golf), circumstances (e.g. wind) or identities (e.g. Austin) as one specific configuration at the physical level. (Dennett, 2004) They are informal predicates. Although vague and subjective, they don't cause any unusual problems. (Dennett & Taylor, 2002) When we say something is possible or could have happened, Dennett (2004; Dennett & Taylor, 2002) says we are talking about a subset of possible worlds that are microscopically different from ours. When Austin, upon missing his putt in a round of golf, says "I could have holed the putt", we should interpret this statement as meaning there exists a subset of at least one similar world (very like our world, but not quite) in which Austin (the entity with the same identity predicate) does hole the putt. If the similarity of possible words is overly limited, we would only be considering identical worlds and thus identical outcomes. If the differences are too strong (say we take a different golf course or an Austin with some additional years of practicing), it would deviate from the initial meaning, but somewhere in between (say we drop the wind-rate by just a fraction or take an Austin who didn't just get distracted by a sinister thought) works just fine. (Dennett & Taylor, 2002)

One of the important features of a relevant subset of CHDO is that its (relevant) informal predicates are obtainable in *this* world. What we need is a way to bridge the actual world into the possible worlds. A person chained to a desk could leave in possible worlds where he isn't (Fischer, 2005), so what is the relevant link to the actual world to be allowed to hold the person responsible for it? My answer to this lies in the efficacy of a moral strategy. We can also put this differently and say that a moral strategy is justified because it itself a condition for a desired otherwise. The thing counterfactuals point to are which dispositions can be reached and in what manner (by which strategy).

Some people, like Jerry Coyne and (to some extent) Harris, object to counterfactual freedom and call it flimsy because it is scientifically untestable. (Harris, 2012) However, I don't believe that is an entirely fair assessment, as science deals with counterfactuals all the time. Without counterfactuals, we couldn't even say that a car can reach 50 km/h unless it is driving 50km/h at the moment we say it. (Dennett, 2014) Nonetheless, counterfactuals are admittedly abstract and difficult to chart. What we need is something that we can treat as evidence for counterfactual competence or excusing circumstances. I don't have any sure-proof suggestions, but I also don't believe it is entirely necessary, to lay out a demarcation-criteria before I am justified in saying we can distinguish, for instance, having missed a putt for a lack of trying and because someone has messed with the golf-clubs.

3.2 CHDO for Free Will

With counterfactuals, we have a meaningful interpretation of CHDO, but further requirements must be satisfied before we consider a can as a freely willed can. Among the relevantly similar conditions for CHDO, there is only a limited set of conditions that justify a responsibility strategy and (hence) classifying the subject as freely willed. The relevant subset of counterfactual, but obtainable conditionals for free will is, I will argue, *reasons-responsiveness*.

Something most accounts of free will have in common is a role, be it central or peripheral, for human reason. An agent isn't freely willed or morally responsible if her actions weren't either directly or indirectly subject to deliberate(d) choices. A deliberate choice is a one which is preceded by a chain of reasoning, a process of deliberation or, quite simply, a *reason*. Not just any kind of reason (a bird catches its prey for reasons of survival), but an "explicit" reason that requires a higher-level intentional stance. Deliberate choices and "explicit" reasons (though they may be based on desires, values or reasoning alone) are the features of a deliberator, a higher-intentional system, "capable of framing beliefs about its own beliefs, desires about its desires, beliefs about its fears about its thoughts about its hopes, and so on." (Dennett, 1997, p. 354) A higher-intentional system is "equipped to notice-and analyze, criticize, analyze, and manipulate- the fundamental parameters that determine its policies of heuristic search or evaluation" (Dennett, 1997, p. 354) in its quest to achieve what is of value. Based on the information about the world and the agent itself in it, it is possible to (though not perfectly) anticipate and assess the consequences of future courses of (in)actions. Essentially, the agent actively tries to act so as to secure a desired future through its own knowledge and evaluations.[1]

Because a deliberator looks towards the future and assesses its courses of action, an appropriate moral strategy will address this capacity. The choices of humans are distinguishable from (most) animal choices by their remarkable sensitive to even minor revisions of societal norms. (Dennett, 2004) This, I believe, is what makes the responsibility strategy such an effective and distinct strategy peculiar to humans. Since there is, to some extent, a relation (or mesh) between an agent's reason and actions, deliberations are a stable way to estimate future actions (Harris, 2012) and thus the most effective (least harmful) focus of judgement. Furthermore, deliberations or reasons are shareable, which opens up the possibility of a responsibility strategy that can give responsibility (share an obligation) on the basis of its rationale (anticipated consequences of value) and/or the promise of blame and praise alone. Agents are excused if there is no violation of obligation and exempted if there is no general capacity to understand or comply with the obligations. (Coates & McKenna, 2015b) It would not be appropriate or justified (meaning, under my account, effective) to blame an agent that lacks the power to recognize and act on the moral reasons supporting the obligation. (Haji, 2002)

Reason is effective as capacity for freedom (attaining what is of value). Unfortunately that relation is not so straightforward. If it were, a mesh between the perfect light of reason and behaviour would suffice. But reason is a tool. It matters very much how you use it and how you are able to use it rather than just whether you used it at all. *Reasons-responsiveness*, a term by Fischer and Ravizza, is a form of control where a sensitivity to reasons aids in acquiring what is of value. (Russell, 2002) Fischer and Ravizza call an agent free if she is receptive to some range of rational considerations (arising from and subject to evaluation, adjustment and monitoring) and able to react or respond to those considerations in the control of her conduct. (Coates & McKenna, 2015a, 2015b) Thus an agent is responsible if the action comes from the reasons-responsiveness of her own mechanism. (Fischer & Ravizza, 2000) I shall elaborate on "own mechanism" later.

Fischer and Ravizza (2000) also address the relation between reasons-responsiveness and responsibility. Too strong a responsiveness to reason would be too tight a fit for morality whereas too weak a reasons-responsiveness would be too loose. There must be an appropriate pattern and regularity in the reasons-responsive mechanism. (Fischer & Ravizza, 2000; Russell, 2002) An agent needn't respond to all good reasons to act otherwise (Matilda might not stop

[1] This includes, incidentally, actions that improve the agent's anticipations and assessments.

dancing for even €100), but does need to respond to a rich pattern of reasons in a coherent (Matilda shouldn't arbitrarily change her policy on when she'll stop dancing), rational (Matilda should only stop dancing for a certain amount of money or more) and sane (Matilda shouldn't only stop dancing for precisely €92,35) pattern. (Coates & McKenna, 2015b) An agent is moderately reasons-responsive if it is regularly, and with an understandable pattern, receptive to a range of reasons and (at least) weakly reactive to those reasons. These points bear some separation in my account. First of all, these are to be approached as considerations for moral strategies. What they point out is that there are ranges to which the responsibility strategy will be effective and/or costly depending on what reasons the agent will respond to. Secondly, it is important that the strategy does no more harm than it saves or than is appropriate. Being too demanding, for instance, might not be sufficiently more helpful and even counterproductive than being a little demanding. Thirdly, due to a lack of counterfactual evidence in every isolated case, it is only if the agent's range of (possible) responses constitutes a coherent, rational and sane pattern that can we justify a modest generalization of the agent's ability to be responsible to similar cases. Fischer and Ravizza (2000) also focus heavily on regularities of range of responsiveness in the actual sequence of history, whereas I think it is more important to emphasize counterfactuals and see regularities or patterns as evidence of those. I believe the counterfactual analysis highlights both the purpose of finding these regularities as well as incorporates the need for evidence of responsiveness that has or could not yet have manifested.

For free will and the responsibility strategy specifically, the relevant subset of counterfactual, but obtainable conditionals is the capacity for more explicit, recursive anticipation and deliberation as the cause of conduct. But only to the extent that counterfactuals highlight that a responsibility strategy can make use of that capacity.

4. Source Origination

We are now ready to move on to the second thesis of free will, which demands that the origin of a choice be with the agent and is relevantly theirs.

4.1 Ultimacy

Under determinism, every action has a cause. If we want to dole out responsibility, the buck must stop somewhere along the chain of causes. The *Source Incompatibilist Argument* states that a person acts of her own free will only if she is the ultimate source of her actions, which, under determinism, is impossible since we can regress the causes (and thus the buck) back infinitely up till the beginning of the universe. (Coates & McKenna, 2015a) The argument relies on the notion of *Ultimate Responsibility*, a term coined by Robert Kane, which asserts that an agent is ultimately responsible for what she did (when she could have done otherwise) if she is the sufficient cause of that action or choice. (Dennett & Taylor, 2002) What fuels the demand for ultimate responsibility is the promise of an *essentialist* regress-stopper.

Nothing forces our hand to accept only essentialist regress-stoppers. We all need a little room for practice to be a self, to form it through continuous self-evaluation. Actions that, after immediate self-evaluation, are subject to systematic revision cannot relevantly be said to contain the agent's intentions. Gradually, less and less of an agent's actions will have passed through without self-evaluation. The agent's actions will become *her* actions because a mesh will unfold itself that enables a reliable assessment of the agent's actions with her intentions.

So, the gradualist can also subscribe to a notion of self-forming acts[2], gradually making up a person who is self-formed enough to not need to trace the cause of her actions beyond her. Even snap judgements won't necessarily be without deliberative force. They can be "remarkably sensitive to myriad features of [the agent's] world that have conspired over time to create [her] current dispositional state." (Dennett, 2004, p. 116) The more a self is formed, the less deliberation it'll require for its purposes and the more (and quicker) it can reveal a self whose decisions are her decisions, making her a proper target for a moral strategy. Causal inquiry is about causally necessary macroscopic conditions, not just the causes which have no further causes. Moral inquiry is interested in which of these causes is an effective target to employ a strategy on, not just the first cause. There is thus no need for one essential cause to stop an infinite regress.

The responsibility bucks stop with the causes that are reasons-responsive, meaning the deliberate choice (or its avoidance) of an agent for which the relevant consequences can be predicted. Each and every one of the deliberate choices made that can reasonably anticipate consequences is subject to a responsibility strategy. An important point that might get lost in the imagery of buck-stopping is that responsibility bucks aren't singular, stopping at the first link in the chain or even shared across the responsible causes. Responsibility is attributed to all the causes that are effectively susceptible to the responsibility strategy. If Lee Harvey Oswald tells a friend of his plan to shoot the president and his friend aids him by giving him a gun to do it, then that friend is not sharing part of the responsibility, but getting his own separate responsibility. So several bucks might stop at several places and they're about as large a buck as their stopping someplace is effective as a moral strategy. Usually, that means they'll gradually die out in importance because anticipating long-term consequences are quite hard and agents need to be self-formed enough to have their deliberations be theirs. The responsibility strategy is used or distributed among the chain of causal necessity as far back as the responsibility strategy is useful.

4.2 Boundaries of an Agent

There is one last concern I wish to deal with and that's how to demarcate the proper target of the responsibility strategy. If a little homunculus or neurological device infiltrated your brain and started to pull your muscles (or your reasons) in a certain way, it would be both physically and functionally inside your system and contribute to your dispositions. Yet we wouldn't call them yours. The reason why, I think, is because there's reason for dividing two mechanisms and assessing them separately.

How we divide into mechanisms will be a pragmatic concern. If a neurological device cannot be detached from the rest of the brain, then there is no pragmatic reason to divide the two. Generally, we'll assess people's choices as the output of their one and only mechanism (the brain), though I think there are exceptions; the neurological device or tumours for instance. If a newly-grown tumour stands in the way of being reasons-responsive, it makes sense to divide the brain into two mechanisms and address the one that's causing problems. If there is no imaginable way to relevantly separate the two physically, we wouldn't say there was any asymmetrical moral considerations about responsibility because the reasons of the composited

[2] Another term by Kane, but used by him to describe an undetermined choice amongst a determined agent's dilemma. ("Robert Hilary Kane", n.d.)

mechanisms were inseparably “hers”. If we can physically (and morally) separate the parts, we (at least hypothetically) do so, if we can’t, we don’t - both in demarcating others as agents and ourself (or our self). Conversely, it is permissible to unify into mechanisms when it makes sense to. You might speak of several different mechanisms in your brain working together to produce your actions, but in most cases, there’s absolutely no theoretical need or practical possibility to separate them. As far as we care then, they are a reliable part of the entity because they can’t or won’t be separated. If mechanisms work together to create a more reasons-responsive entity with a richer or better variety of alternate possibilities, then the division becomes unnecessary in theory and morally wrong in practice. They compose one larger mechanism, one agent.

The point about ownership has to do with finding the proper target of a moral strategy. Once we have found the target(s), we can assess its dispositions and role in the causal chain to pick a moral strategy as a response. When we excuse people from responsibility because they were manipulated or mediated, it is only to the extent that their reasons-responsiveness was either inhibited by another mechanism, altogether absent or insufficiently formed to be reliably considered as their stamp on the world. However, as long as there is a reliable connection with other mechanisms, they may be combined (e.g. the brain). In that case it is the combination, as one agent that is reasons-responsive and the combination that is justifiably held responsible to the extent that the combination can now respond to the responsibility strategy. If, conversely, the part(s) are freer than their whole, we divide them so that a moral strategy can attempt to target its parts instead.

6. Artificial Source Origination

I have now given a compatibilist reading of the both theses of free will that set up a free will worth wanting. In principle, there appear to me to be no worthwhile objections to an artificial agent satisfying these conditions, meaning that the same points about free will likewise apply to the appropriate AI. Allow me to, in broad terms, apply it. To begin, let’s consider the second thesis of source origination in an artificial context. This requirement demanded that the origin of a choice or intention lies inside the agent. With Artificial Intelligence it seems reasonable to ask: who do the intentions belong to?

When can we say something or someone has intentions? Dennett (2009) has given a clear answer to this question. Attributing a person or a system with an intention is part of taking an *intentional stance*: utilizing a useful level of abstraction (with its own ontological categories) to explain certain kinds of behaviour. From this stance, we treat the designated entity as an agent with beliefs, intentions and enough rationality to do what it was designed to of those beliefs and desires. The concepts of this stance (beliefs, desires, rationality) make its behaviour both predictable and explanatory. (Dennett, 1997) A thermostat could be said to be an agent, since it is fruitful to speak of its actions in terms of a belief about the room temperature and a desire to achieve a different one. Admittedly, in the case of thermostats such a fanciful anthropomorphism is a bit unnecessary. However, with systems of sufficient complexity an intentional stance becomes indispensable (from our human perspective) for successful interaction. Taking an intentional stance towards our fellow creatures is what we do each time we attribute intentions, beliefs to what we consider to be (limitedly) rational agents. (Dennett, 2009)

Dennett goes on to make the crucial argument that the difference between an intentional system and any other system is not a special metaphysical tag that distinguishes "real" intentions from imagined ones. "To seek a clean distinction between some metaphysically authentic intentional beings and simulacra like thermostats presupposes that there is more to any intentional system than adopting a stance toward it as an intentional system." (Coates & McKenna, 2015a) Deciding which stance to take depends and is justified solely by its practical efficacy.

Lady Ada Lovelace wrote of computers that they "ha[ve] no pretension to originate anything. It can do whatever we know how to order it to perform" (cited in Turing, 1950/1985). Turing (1950/1985) reframed the objection to whether a machine can "take us by surprise" and this they most certainly can. Nonetheless, the critic insists that the surprise reflects no credit to the machine, but to the creativity of our own minds. Machines can't, in principle, surprise us because they can't do anything "really new". (Oppy & Dowe, 2011; Turing, 1950/1985) Bringsjord (2001) defends Lovelace on the same ground. Software doesn't always do what it is intended or what we expect it to, but only - he insists - because we're not smart enough. He says a computer program can only take it "upon itself to originate something" (p. 5) if it can reliable repeat an action (to rule out malfunctions) for which it was not programmed and which cannot be explained by even the designer by appeal to the program's "architecture, knowledge-base, and core functions" (p. 12). The reader will be reminded of the source origination objections to free will. Winograd & Flores (1990) likewise defend the notion that a computer cannot originate anything. They argue that any program is a mere intermediary medium to the commitment (intentions) and responsibility of the programmer. This objection is analogous to the infinite regress of causation that troubled us with our own intentions. Here it is assumed that humans stop the regress of commitment, but they never do clear up how humans originate any such commitments. Only if one starts from the premise that humans are not determined, does this argument hold any weight. Since I have built my account of free will on the basis of determinism, this argument simply doesn't apply.

As I have argued before, the infinite regress can be stopped. Not through a single unpredictable act over which we have no meaningful control, but through the gradual acquirement of the, necessary causal, reasons that are internal and stable enough to warrant a reliable intentional stance towards the system. Both creator and creation play their own causal role in the events. Dennett (1997) clarifies where the originative credit is due by example of the computer chess-program *Deep Blue*. Its behaviour is sufficiently complex and systematically purposeful to warrant an intentional stance. When Deep Blue beat world chess champion Garry Kasparov, it was "Deep Blue's sensitivity to those purposes and a cognitive capacity to recognize and exploit a subtle flaw in Kasparov's game that explain Deep Blue's success" (Dennett, 1997, p. 352) and not the designer's. The designers do, of course, play a necessary causal and intentional role in Deep Blue's cognitive capacity. However, congratulating the designers is much like congratulating the educators or parents of (or the process of natural selection leading to) Kasparov, his human opponent. (Dennett, 1997) Though that is certainly justified, it doesn't take away Kasparov's credit as a world chess champion. The sequence of winning moves was found by Deep Blue. So it were Deep Blue's reasons and Deep Blue's intentions that are a (though not the only) proper target of using an intentional stance. If Deep Blue was capable of appreciating praise, we would justifiably deploy that strategy. Deep Blue's capacities are not wide enough to qualify as morally responsible, but they are strong enough to qualify as the originative winner of the match.

For an action to be up to a system, the necessary cause needs to have been its internal programming (psychology), regardless of how it came to be that way. As soon as the intentional stance proves its effectiveness, a moral strategy will make use of it and regard the intentions as those of the system's rather than mediate it to another (though it might do that as well). With regard to the responsibility strategy, distinguishing responsibility from mediated responsibility requires nothing other than where the reasons-responsive buck stops (and doesn't).

7. Artificial CHDO

A computer system can be the relevant causal source of an event, but that doesn't give it free will yet. The first thesis of free will required that the agent possessing free will has the ability to could have done otherwise, and for *reasons*. I will argue that, provided a system has some capacity to learn, they have a range of otherwises at their disposal.

7.1 Artificial Otherwises

To consider the importance of learning, I'd like to address another objection to AI: *the Argument of Informality of Behaviour*. Turing (1980/1985) described the argument of informality of behaviour thusly: "It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances." (p. 65) Any set of circumstances without the proper rules for responding would result in errors. To retort, Turing points to a confusion between "rules of conduct", stipulating precepts which can be followed consciously (e.g. stopping for a red light) and "laws of behaviour", which are laws that regulate our bodies (including our brains). It is the latter that are of importance to produce AI and, since we are at least near-determined, there is insufficient cause to say there are no such laws for human beings. (Turing, 1950/1985)

I would personally like to add to Turing's remarks that the argument of informality of behaviour misconstrues how intelligence works. I could see how having a ready rule for every exact situation could allow for the most intelligent response in each. However, not only would that be a ludicrous demand, it doesn't solve how it was determined to be the most intelligent response. On top of that, it doesn't correspond to what we call intelligence. It is the mark of intelligence that it can operate even under circumstances for which there is no ready solution or stipulated rule of conduct. While it is true that some behaviour and even elements of representation can be wired-in, others must be learned. (Dennett, 1993) A certain amount of plasticity should be admitted to allow the system the room for trial and error, in actuality and deliberation (imagined anticipation), to deal with novel situations. However, by what system can intelligent responses be evaluated *to* learn?

There is a prevalent evaluation system already in place in the world. Agents who lack self-persistence and whose interaction with the environment are less than ideal are ill favoured by natural selection. The (implicit) reward of receiving a non-negative evaluation by natural selection is to live and to reproduce. Perdijk (2014) calls this generational evaluation because the evaluation can only occur by terminating the agent or obstructing reproduction. Mutations that provide beneficial behaviour can be considered, in a broad sense of the word, to be learned across generations. It is not the agent, but the species that learns through natural selection. With regards to moral strategies, there seems little room beside isolation or destruction in this case.

A degree of plasticity in the brain with an (even limited amount) of capacity to be punished or rewarded allows agents to learn traits in their own lifetime (we called this post-natal design fixing). For that we need an evaluative reward mechanism inside the agent. The active role of pain and pleasure, regardless of discussions regarding their subjective force, are obvious here. (Dennett, 1993) For now, we haven't yet strayed beyond an artificial reactive agent with some parameter-adjustment of set rules for behaviour that correspond to set rules of conduct. Better agents would be able to redesign themselves depending on the conditions they encounter. (Dennett, 1993) This is the case with state-determined agents, whose actions can also be driven by remembered (or, with sufficient complexity, expected) reward and punishment. (Perdijk, 2014) Natural selection is now at work at both the level of the species and at the level of the agent, the latter of which Perdijk (2014) calls lifetime evaluation. In this case, it can learn through non-fatal encounters that are still subject to actual or anticipated punishment or reward. With a sufficiently complex state-determined agent, anticipations of complex compositions of expectable situations it hasn't (yet) encountered can prepare an agent's responses.

Once flexible state-determined systems are allowed, we may consider CHDO. Dennett (2004; Dennett & Taylor, 2002) uses the example of a chess program to illustrate it. Let's say the pivotal move in a game of the program, the necessary cause of its downfall, was a failure to castle. It would be meaningful to ask whether the program could have castled or not. Castling is a legal move, sure, but the question is whether it was within the program's dispositional repertoire to castle. Our chess program is determined, so looking at the conditions exactly as they were would be uninformative because its actual outcome in those precise circumstances will have always been the same. However, wiggling the events will show whether it was a remote, big or no possibility for the program. If we find that a relevant subset of similar conditions the chess program does castle, then there is a meaningful sense in which we say the program could have castled. (Dennett & Taylor, 2002) It is in that sense that we assess the program's capacity to play chess.

7.2 Artificial Reasons-Responsiveness

Learning is not only necessary for intelligence, but for free will as well. An agent can only respond to a responsibility strategy if it can act, or learn to act, otherwise in accordance with a moral rationale (or moral attitude) that is transmissible. This is what I have earlier called the right kind of CHDO.

Computers, as we currently know them, have their hardware designed to start out empty and accept universal software. So the interchangeability of software in current computer systems allows for horizontal learning. And indeed, if all that stands between an Android murdering and not murdering is a software-update, such a moral strategy would be effective. Nevertheless, such a strategy is not quite a responsibility strategy. Though there is a similarity between the mind's software and that of computers, our minds do not start out empty. Nor is our software as universal or efficiently shareable as that of computers. It is, however, much easier to work with in another sense. Computer-software has to be strenuously programmed before it can be shared. Programming languages make this a great deal easier and more intuitive, but not quite as effective as communication between two ordinary human adults. Communication (and instruction) between two human adults can be incredibly fast, flexible and pervasive in sharing its content because it is subject to a recursive (self-)control structure as well as a high amount of software-interconnectedness to maximize the effect. One meme can change someone's entire mind. A single piece of software has no such (functionally efficient) effect on other software. Part of the problem, I believe, is answered by having a recursive self-control structure. A single piece of information can be broadcasted (made famous) and looped around to great effect. However, to be able to do that, software needs to also be highly interconnected. If a piece of recursive reasoning can't affect anything beyond reasoning, it wouldn't do much good. The rationale must be captured by the whole (or relevant parts) of the system.

To recognize a rationale (in most cases) requires a complex assessment of the world and one's own actions in it.[3] A freely willed AI should not only be able to make decisions that it leaves to natural selection to evaluate, but be able to evaluate those decisions itself through complex anticipation that is itself subject to evaluation. It should be able to take an intentional stance towards itself. It should treat itself as a moral agent, evaluate its own behaviour and its own selection of behaviour. A self-symbol, sometimes suggested as a requirement for self-consciousness, would not suffice to acquire a virtual captain, because the self needs to be sophisticated (a system rather than a single symbol) and functional enough to be utilized as such. (Hofstadter, 1999) In short, it needs to be a higher-intentional system. As a higher-intentional system, it can make predictions about the environment, its own behaviour and its own predictions. And most importantly, it can then evaluate each of those for accuracy and desirability, broaden its epistemic horizon and recognize the rationales (positive consequences) of complex actions. This is why moral strategies focus heavily on training and holding people responsible to stretch the epistemic horizon of their choices, to take responsibility. Whatever intentions (and consequences) cannot (reasonably) be captured by the self-narrative cannot be a target of a responsibility strategy since its scope only stretches the length of obtainable otherwises that are within the (reasonable) epistemic horizon of the system.

Capturing a rationale, of the kind required for free will, requires not only its recognition, but that it can be moved or taught via the rational persuasion of either others or of itself. Essentially what we need is a recursive (self-)control structure. If a system can make reliable use of a rationale through its recursive self-control, then we are permitted to say it has captured it. Then its anticipations of its own actions and consequences can be used to better effectuate its own behaviour by having reasons (subject to revision) for its actions. Through this process, the higher-intentional system can form a better self. To the extent that it can do that, it can respond to the challenges of moral responsibility. At least, if there is a form of communication that can access the agent's self. Communication may be in whatever language. It may be visual,

[3] There is an interesting elucidation of reasons-responsiveness by linking it with leading theories of consciousness such as Baars' (2001) Global Workspace Theory, but the elucidation of this connection is beyond the need and scope of this paper.

auditory, or whatever sense an agent has at its disposal as long as it does the trick of spreading the rationale. What is of crucial importance to a responsibility strategy is the ability to share the anticipations, evaluations and rationales that have an effect on how the system anticipates, evaluates and teaches itself. If it can do that, the system should be able to actually (or limitedly counterfactually) respond to instructions by being able to recognize and respond to a rationale behind an instruction, be reasons-responsive, and thus a fit subject to employ a responsibility strategy to.

8. Kludged Free Will

If we want to assess an AI with free will, we do have to make sure that what we are seeing is evidence of the machine's sophisticated mechanism of recursive control, not the illusion of it offered by kludges. A kludge is computer jargon for a trick, a short-cut for computers to appear cleverer or more complex than they actually are. (Sharples et al, 1994) Examples of kludges are canned expressions of emotional responses that get attached at appropriate moments. However, as Dennett (1997) points out, "many of our own avowals of emotion are like that - insincere moments of socially lubricating ceremony" (p. 361) Therefore, I believe kludge to be a useful term for humans, animals and computers alike. We are very susceptible to attributing complex inner lives whenever we see the opportunity for it. There is a now famous example of a vice president of a computer company stumbling upon a version of the kludge-program ELIZA and mistaking it for a Teletype connected to an employee's residence, whom he believed to be deliberately evasive to the crucial questions put to him. (Güzeldere & Franchi, 1995) It usually takes some extra effort to realize nobody's really home. Nonetheless, many of these kludges, exactly because they are limited to short-cuts, can be found out by their lack of sophistication or (reason-)responsiveness.

I believe it is kludges most specifically that are of primary concern in the scepticism of artificial intelligence. The concern, I believe, is that all AI will be little more than a bag of kludges, sufficient to give the appearance of conscious intelligence, but insufficient in its devices to replicate the actual property. This certainly appears to be a justified scepticism with regards to current AI, but I don't believe this concern should be made inherent to its endeavours. Since we are more familiar with the evidence for real or kludged reason with humans, they are sometimes easier to find out than those of artificial intelligent systems - which may be designed in numerous ways. However, a kludge will always betray itself in performance or it wouldn't be a kludge. Something is not a kludge just because it is "programmed that way", but because it appears to be (programmed for) doing something more complex than it is actually doing. Should the devices succeed in replicating sufficient complexity of consciousness, then it is the property, and not just the appearance, that it has replicated. A sophisticated recursive (self)control structure, capable of capturing, receiving and executing rationales would constitute reasons-responsiveness and thus allow the counterfactual assessment of abilities which we use to determine free will.

9. Conclusion

I have argued that free will is a property attributed to an agent which can respond to a specific utilitarian moral strategy: the responsibility strategy. To employ a utilitarian moral strategy, we must begin by considering whether an agent is a proper cause for an action that could have been otherwise. Evidence for this benefits from a counterfactual analysis of the agent's dispositions, making for a meaningful interpretation of "could have done otherwise". If the difference causing the otherwise is reasons-responsiveness, then this points to the efficacy of the responsibility strategy and thus the agent's free will. The internal psychology or programming of an agent, regardless of how it acquired it, makes a system a reliable target to attribute the intentions subject to a moral strategy. A system capable of learning opens up a range of obtainable otherwises. Being able to learn behaviour via horizontally shared rationales, acquired by others or itself makes a responsibility strategy effective. At least, if there is enough interconnectedness of software for rationales to have a responding effect. Evidence of this counterfactual ability will be the right kind of "could have done otherwise" for responsibility.

Reference List

Baars, B . J. (2001). *In The Theatre of Consciousness: The Workspace of the Mind.* New York: Oxford University Press.

Bringsjord, S. (2001). Creativity, the Turing Test, and the (Better) Lovelace Test. *Minds and Machines*, 11, 3-27. Retrieved from http://kryten.mm.rpi.edu/lovelace.pdf

Coates, D. J., & McKenna, M. (2015a). Compatibilism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2015 ed.).* Retrieved from http://plato.stanford.edu/entries/compatibilism/ (22/09/15)

Coates, D. J., & McKenna, M. (2015b). Compatibilism: State of the Art. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2015 ed*.). Retrieved from http://plato.stanford.edu/entries/compatibilism/supplement.html (22/09/15)

Dennett, D. C. (1997). When Hal Kills, Who's to Blame? Computer Ethics. In D. G. Stork (Ed.), *Hal's Legacy: 2001's Computer as Dream and Reality (*pp. 351-365). Cambridge: MIT Press.

Dennett, D. C. (2004). *Freedom Evolves*. London: Penguin Books.

Dennett, D. C. (2009). *Intentional Systems Theory.* (n.p.). Retrieved from http://www.lscp.net/persons/dupoux/teaching/QUINZAINE_RENTREE_CogMaster_2011-12/Bloc_philo/Dennet_2009_intentional_systems.pdf

Dennett, D. C. (2014). *Reflections on FREE WILL.* [Review of the book Free Will, by S. Harris]. Retrieved from http://www.samharris.org/blog/item/reflections-on-free-will

Dennett, D. C., & Taylor, C. (2002). Who's Afraid of Determinism? Rethinking Causes and Possibilities. In R. Kane (Ed.), *The Oxford Handbook of Free Will* (pp. 257-277). New York: Oxford University Press.

Fischer, J. M. (2005). Reply: The Free Will Revolution. *Philosophical Explorations*, 8(2), 145-156.

Fischer, J. M., & Ravizza, M. (2000). Précis of Responsibility and Control: A Theory of Moral Responsibility. *Philosophy and Phenomenological Research*, 61(2), 441-445. Retrieved from http://www.jstor.org/stable/2653660

Franklin, R. L. (1968). *Freewill and Determinism: a Study of Rival Conceptions of Man*. London: Routledge and Kegan.

Güzeldere, G., & Franchi, S. (1995). Dialogues with Colorful "Personalities" of Early AI. *Stanford Humanities Review*, 4(2), 161-169. Retrieved from http://www.stanford.edu/group/SHR/4-2/text/dialogues.html#note10

Haji, I. (2002). Compatibilist Views of Freedom and Responsibility. In R. Kane (Ed.), *The Oxford Handbook of Free Will* (pp. 202-228). New York: Oxford University Press.

Harris, S. (2012). *Free WIll.* New York: Free Press.

Harris, S. (2014). *The Marionette’s Lament: A Response to Daniel Dennett* [Blogpost]. Retrieved from http://www.samharris.org/blog/item/the-marionettes-lament

Hofstadter, D. R. (1999). *Gödel, Escher, Bach: An Eternal Golden Braid.* New York: Basic Books.

Oppy, G., & Dowe, D. (2011). The Turing Test. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Spring 2011 ed.).* Retrieved from http://plato.stanford.edu/archives/spr2011/entries/turing-test/

Perdijk, N. (2014). *Artificial Reward and Punishment: Grounding Artificial Intelligence Through Motivated Learning Inspired by Biology and the Inherent Consequences for the Philosophy of Artificial Intelligence.* MA-Thesis, University of Utrecht.

Russell, P. (2002). Pessimists, Pollyannas, and the New Compatibilism. In R. Kane (Ed.), *The Oxford Handbook of Free Will* (pp. 229-256). New York: Oxford University Press.

Sharples, M. et al (1994). *Computers and Thought: A Practical Introduction to Artificial Intelligence.* Cambridge, MA: MIT Press.

Singer, P. (2011). *The Expanding Circle: Ethics, Evolution, and Moral Progress.* Princeton: Princeton University Press.

Turing. A. (1985). Computing Machinery and Intelligence. In D. C. Dennett & D. R. Hofstadter (Eds.), *The Mind's I: Fantasies and Reflections on Self and Soul* (pp. 53-67). Middlesex, England: Penguin Books. (Original work published 1950).

Winograd, T., & Flores, F. (1990). *Understanding Computers and Cognition: A New Foundation for Design.* Norwood, NJ: Ablex Publishing Corporation.