SINFUL AI

Michael Wilby

When asked what he thought the future held in the wake of his invention, Geoffrey Hinton, the 'Godfather of AI', and one of the originators of the architecture behind generative AI Chatbots such as GPT4, was blunt: 'my intuition is: we're toast. This is the actual end of history'. The threat that Hinton thinks AI poses would seem to be the epitome of what the concept of evil was designed to describe: an intelligent but alien being with the means and motivation to turn the whole world to dust; an artificial Satan or Mephistopheles. AI futurists have not been slow to come up with scenarios where AI runs out of control, either deliberately or accidently. For instance, the Swedish philosopher Nick Bostrom devised the infamous example of a 'paperclip maximiser'. This is an AI that is designed to produce as many paperclips as it can, and ends up turning everything in the world, including living beings, into paperclips.

Nevertheless, AI-apocalypse scenarios really are just ways of expressing an underlying fear. Like the Sorcerer's Apprentice, we seem to have created something that we don't fully understand, and the consequences of which are potentially far-reaching and destructive to our way of life. We know that AI is capable of great harm, but it is not clear that we have the necessary concepts and vocabulary to assess these threats in moral terms. It seems doubly wrong, for instance, to describe the threat of artificial systems making us 'toast' as merely a bad outcome. It seems to be something much more than that. But what? Can we make sense of AI as being evil? What would this mean?

It is essential that we make sense of what it could mean to say that AI could be held responsible not just for *moral wrongdoing*, but also for *extreme moral evil.* To do this, we will have to take the scenic route. It won't be enough to just point to the potentiality of a Terminator-style robot and label it 'evil': we need first to get a grip on what it would mean to call

something 'evil' and then get a grip on how such a concept could sensibly be applied to a non-living machine.

Taxonomy of the Threats

There are various ways in which we might label the risks that AI poses, depending on whether we focus on the harm caused or on the causes of the harm. If the former, then we might turn our attention to AI risks such as surveillance, bioterrorism, automated warfare, technological unemployment, the control problem, and value alignment. However, given that the focus here will be on AI and the concept of evil – that is, trying to understand the extent to which an AI can be morally responsible for the outcomes connected to it, including where those outcomes are extreme and our ordinary moral responses start to run out – we shall instead discuss, at least in the first instance, the causes of the potential harm. This will require looking at how AI can admit of differing degrees of agenthood which, in turn, might map onto differing degrees of moral responsibility. Here is a fairly crude, but hopefully useful, taxonomy of three forms of AI agenthood:

(a) AI as a *non-agential tool*
(b) AI as a *hybrid, minimal agent*
(c) AI as a *fully autonomous agent*

To what degree would it be correct to think of any of these forms of AI agency as incorporating moral responsibility? Can a mere tool – even a highly advanced tool – be held morally responsible? Should the deadly effects of an out-of-control automated car be the responsibility of the manufacturer, the on-the-loop driver, or the car itself? Further, to what extent would it be correct to say that such systems are not just capable of wrongdoing, but are capable of evil, in the narrow sense to be explained below?

One way of framing the issues of AI moral responsibility is in terms of responsibility gaps. Sven Nyholm, a professor of AI ethics at the Ludwig Maximilian University of Munich in Germany, characterises responsibility gaps as follows:

> when we start using technology that take over tasks from human beings, like robots and other AI technologies do, there are often worries that there might be cases when the stakes are high and when outcomes might come about for which somebody should intuitively speaking be held responsible. But it might be unclear who, if anybody, could or should be held responsible. So potential responsibility gaps occur.

The concept of a responsibility gap suggests that an artificial system can be thought, in moral terms, as being no more than a tool. It suggests that moral responsibility, if it is to come, must come from the human operator; and, where the human operator has ceded control to the machine, then we are left with a moral vacuum – a responsibility gap that is generated by the (supposed) fact that artificial machines are not susceptible to genuine moral censure.

We need to scope out the extent to which, in contrast to the assumption governing the responsibility gap, AI itself can be held morally responsible for certain acts, in situations where we standardly would hold agents morally responsible, and then to consider how and when (and if at all) the concept of 'evil' specifically can be applied to AI. Scepticism about applying moral concepts to machines and AI is often unconsciously conflated with a wider scepticism about agency and morally responsibility in general. It is often assumed, for instance, that AI cannot be held responsible because they lack [...]. What fills the gap here could be various (consciousness, intentionality, free will, and so). Such concerns often revolve around a sense that AI is physically determined and hence cannot house these properties and be a subject of genuine agency or moral responsibility. However, once we start inquiring into human practices of holding each other responsible, rather than looking for metaphysical criteria of being morally responsible, questions of determinism fall out of the picture, and the way is open to considering whether the full gamut of moral responses, including the concept of evil, can be applied to AI.

Answering the Evil-Sceptic

AI can, within certain circumstances, be held morally responsible. Moreover, we can use the concept of 'evil' to illuminate some of the threats posed by AI. To understand how concepts of moral responsibility can be

applied to machines, however, we will need to go back to the beginning and outline a framework that will provide an account of how so much as *any* agent can be held morally responsible for their acts. This framework can be used, I shall argue, to provide a basis not just for the ordinary run of moral concepts, such as 'good', 'bad', 'right', 'wrong', but also for moral concepts that sit at the extreme, like 'evil'.

There are two forms of scepticism about the concept of evil. A narrow scepticism that questions the reality of evil specifically, and a wide scepticism that questions the reality of morality in general.

The wide form of scepticism about evil argues that evaluative concepts in general are problematic, and so, by extension, is the concept of evil. Scepticism about morality in general has a long history, and numerous variations. There are both descriptive versions of the challenge, as noted by the late Australian philosopher John Leslie Mackie and normative versions of the challenge, as discussed by the nineteenth century German philosopher Friedrich Nietzsche. We can focus on two main points, however (both descriptive): (a) that moral concepts purport to pick out moral facts; but since moral facts would be unusual or 'queer' entities – non-empirical and non-physically located entities – then we can assume that they don't exist (The Argument from Queerness); (b) moral concepts depend on moral responsibility, and moral responsibility depends on a capacity for libertarian free will. But since libertarian free will is also a 'queer' entity – it is an ability to act in an undetermined way – so we can assume that there is no such thing as moral responsibility either (The Argument from Free Will).

The narrow form of scepticism about evil argues that the concept of evil specifically is a problematic concept, even though moral concepts in general are of good standing; it is problematic for two reasons. First, it is descriptively inaccurate and fails to pick out anything that exists in the actual world, as opposed to the world of myth and fiction. Second, it is normatively questionable because it leads to the demonisation and dehumanisation of people. These two points are complementary: because 'evil' purports to pick out inhuman mindsets or intentions, then, once we put the concept under scrutiny, we find both that there are no such mindsets and intentions, and that using the language of evil immediately dehumanises those it is applied to.

These are very brief sketches of arguments that have, of course, been developed in considerably more detail through the history of modern philosophy. To show how these forms of scepticism can be responded to, we can appeal to a framework that draws on the English philosopher P F Strawson's influential work on what he calls the 'reactive attitudes'. What is useful about this framework is that it is specifically designed to address concerns about moral scepticism that derive from the apparent incompatibility of moral responsibility and physical determinism. Concerns that genuine responsibility – and in its turn, genuine evil – requires either a libertarian free will, or queer, supernatural entities are ways of speaking about this clash between the natural and the normative. Since, as I have already intimated, this clash is also partly what drives scepticism about the applicability of moral responsibility to AI, we will, with this framework in hand, be able to return to the question of morally responsibility and the application of 'evil' to AI.

The Reactive Attitudes and the Moral Nexus

The reactive attitudes are, as Strawson puts it, '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'. For instance, if I were to sullenly push you in the back, then that would invoke in you a feeling of blame or resentment, directed towards me, on account of what I have done. In reacting this way (with attitudes of blame and resentment) you would be reacting to something within my behaviour that expresses an attitude of mine of a lack of regard, or a quality of ill-will, towards you. Reactive attitudes such as these constitute something of a semi-articulate normative system. My attitude of ill-will or disregard should invoke in you a corresponding attitude of blame, which – assuming the blame is well-judged – should then invoke in me an attitude of remorse; and then – assuming the remorse is genuine – it should invoke in you an attitude of forgiveness. Where some of these attitudes are not well-placed, then the appropriate responses might be different: justifications, excuses, indignation, and so on. Put together, the reactive attitudes can be understood as elements of a practice that consist in an

interrelated nexus of responses to the attitudes, feelings, and behaviours of others.

What is key to Strawson's approach is a reversal of the standard order of metaethical explanation. Rather than understanding moral reactions as reflective of an underlying set of moral beliefs about what is right or wrong, Strawson instead suggests that we should regard those moral reactions as elements of a practice full participation within which is constitutive of what it is to be morally responsible. In other words, one *is* morally responsible to the extent that it is appropriate, as a member of that practice, to be *held* morally responsible. Insofar as one is not a member of that practice one is not held fully responsible within that practice, and so is not fully responsible by the lights of that practice.

The Reactive Attitudes and Ordinary Moral Practices

Strawson regarded his framework as providing a response to the wide form of scepticism about evil. That is, he suggests that appeal to the reactive attitudes provides a framework for a non-reductive, but naturalistic understanding our ordinary moral practices – practices of responding to each other with attitudes such as blame, praise, gratitude resentment, and so forth – that relies neither on the idea that there is a realm of independently specifiable moral facts, nor on the idea that moral responsibility requires a radical libertarian free will. The framework satisfies both the Argument from Queerness and from Free Will. We can briefly outline how these arguments are meant to go.

An appeal to independently specifiable moral facts is not required because, on the Strawsonian view, moral responsibility can be fixed by criteria internal to the practice. For instance, a basic demand for recognition between participants of the practice calls for responses to harmful behaviour (at least where that behaviour is expressive of ill-regard) such as blame, excuse, or remedy. The practice itself might be without external justification: it might just be that holding each other responsible in these reactive ways is what we (perhaps contingently) do, or, in some circumstances, what we feel psychologically compelled to do.

An appeal to a radical libertarian free will is also not required for moral responsibility in line with the Strawsonian view because participation in

the practice does not require an absolute conception of free will. This is true in both directions of an interaction. From the point-of-view of the blamed, whether one is expressing ill-regard for another is independent of whether it is fully freely chosen; the attitude is one thing, the cause of it another, and the attitude would remain even if it was causally determined. From the point-of-view of the blamer, one's response to another's ill-regard is neither diminished nor extinguished by a philosophical conviction about determinism. That is to say, it is not part of the internal criteria of the practice that one modify one's sense of blame, say, on account of a global conviction about the metaphysics of personal identity; rather, the reactive attitudes are modified in response to particular circumstances (which can generate excuses or justifications for immoral behaviour) or in response to particular facts about a person (for example, on account of being under stress at the time, being ill, or being a child).

So, one modifies or abandons one's reactive attitudes – and hence abandons one's assumption that another is morally responsible or blameworthy for an act – only in particular or 'abnormal' circumstances. This means that we cannot globally abandon claims of moral responsibility, as would be the case if we accepted the wide form of scepticism about evil, because 'it cannot be a consequence of any thesis which is not itself self-contradictory that abnormality is the normal condition'. Again, notes Strawson, 'the human commitment to participation in ordinary interpersonal 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'.

What we have here, then, is a basic framework for understanding how morally responsibility – grounded in a participatory practice in which participants hold each other responsible by way of reactive attitudes – need not involve an appeal to anything beyond ordinary, naturalistically describable properties of the attitudes and actions of the agents involved. There is no need to appeal to libertarian free will, or to spooky non-natural properties of goodness. This, then, opens up – but does not yet establish – an explanation for how AI can be properly held morally responsible. To the extent that an AI is able to engage in the practices that constitute holding and being held responsible, then AI can be thought of as *being* morally responsible for its actions.

The Reactive Attitudes and Marginal Cases

I have argued, then, that the Strawsonian Framework provides a sketch of an account of moral responsibility that promises to exonerate our ordinary everyday moral practices from at least some of the charges outlined by the wide form of scepticism about evil; in particular, the Argument from Queerness and the Argument from Free Will.

However, as observed by various commentators, Strawson's framework has a *prima facie* difficulty placing what the American moral philosopher David Shoemaker calls 'marginal cases'. Marginal cases are 'cases at the boundaries of our interpersonal community where agents tend to strike us as eligible for some responsibility responses but not others'. For instance, a young child who misbehaves will be held responsible for what they do, but not in the same way that an adult would; in some ways and in some respects their responsibility is diminished. This means that a crude picture of the Strawsonian Framework – in which one is either a member of the moral community or one is not – cannot be the whole story; we need the framework to account for marginal cases.

Although Strawson acknowledges 'essentially a borderline, penumbral area' of marginal cases, his framework struggles to accommodate them. This is because Strawson contrasts his framework of *participant reactive attitudes* – 'essentially natural human reactions to the good or ill will or indifference of others towards us' – with what he calls the *objective attitude.* The objective attitude involves seeing another being as, either temporally or permanently, outside the normal range of moral and interpersonal participation – as 'a subject of social policy; as a subject for what, in a wide range of sense, might be called treatment'. Although the objective attitude can be 'emotionally toned' in various ways, it does not 'include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships'.

To get a grip on the distinction between the participant stance, and the objective stance on the other, we can consider two broad ways in which might inhibit one's negative reactive attitudes (such as blame or resentment) towards another.

The first – sometimes called *'Type-1 Pleas'* – are when an excuse or justification can be given for an action. For instance, if I miss an

appointment with you, I might be excused if it turns out that the missed appointment was due to a factor outside my control, for example a late-running train. In cases such as these – or cases where there are small lapses due to ignorance or tiredness – there is no tendency to regard the other as outside the scope of moral responsibility. As Strawson notes, 'exculpatory pleas … in no way detract … from the agent's status as a term of moral relationships'. Type-1 pleas are part-and-parcel of the participant stance of the reactive attitudes.

The second broad ground for suspending or inhibiting one's negative reactive attitudes – sometimes called *'Type-2 Pleas'* – are when an agent is seen as not capable of proper moral engagement with others. Perhaps, for instance, they are suffering from some 'insane delusion' about the world that doesn't enable them to understand what is happening around them, or perhaps their behaviour is not fully under their control, and they are subject to sub-conscious or non-intentional compulsions or behaviours, or perhaps they are simply, as Strawson puts it, a 'moral idiot' who lacks the capacity to empathise with others. These promote 'the purely objective view of the agent as one posing problems simply of intellectual understanding, management, treatment and control', rather than problems of proper regard in interaction; such a person is (perhaps only temporarily) not 'seen as a morally responsible agent, as a term of moral relationships, as a member of the moral community'. Type-2 pleas place an agent *outside* the participant stance of the reactive attitudes.

There is a need, then, to understand the Type-2 marginal cases, the 'ever-interesting cases of variation' as Strawson called them. There has been significant progress in this direction. Shoemaker, for instance, has argued that we can understand our moral practices as involving three distinct ways of being held responsibile – attributability, answerability, and accountability – that correspond to three distinct ways in which a person might express a poor quality of will. On this view, an act is attributable to an agent if the act is reflective of, and caused by, that agent's deeply held cares and commitments (American philosopher Susan Wolf's 'deep self'). An agent is *answerable* for an act when that act could be regarded – either hypothetically or actually – as being the outcome of a deliberative choice that considered relevant alternatives to the action taken. And an agent is

accountable for an act when they are a fitting subject of participatory reactive responses on account of their quality of regard.

Now, Shoemaker argues that by understanding moral practices in this way – as involving three forms of being held morally responsible – we can understand marginal cases as involving the invocations of some forms of responsibility without others. So, while a fully-fledged adult member of the moral community will generally be held responsible for their actions and attitudes in all three ways, more marginal members might be exempt from being held responsible in one or more ways. For instance, a child might be attributively responsible or even answerable for their behaviour in certain respects, but they might not be accountable, because they are not capable of properly holding another in the right regard, even though the action they performed was both in line with their central cares and commitments and was committed after some deliberation.

Likewise, in a way that speaks to our current topic, Shoemaker's account could help us understand the sense in which Artificial Intelligence – either current or soon to come – might be held responsible in a minimal way: to the extent that AI is capable of being sensitive to harm against others, and capable of deliberation about means, but not capable of having its own ends. Then we might hold AIs accountable and answerable for their actions, but not attributable, at least in the sense outlined by Shoemaker.

The Secular Problem of Evil

However, marginal cases do not just include agents with what we might call 'minimal' capacities, such as young children or AI. It also includes agents who commit what we would colloquially call 'evil acts'. Such acts invoke in us a contradictory response: extreme offenders at once seem incapable of engaging in ordinary practices of moral accountability – and to that extent seem to fall outside of the practice, participation within which, we have argued, is necessary for moral responsibility – while also eliciting, among ordinary members of the moral community, a sense of undeniable moral outrage and moral blame for their actions.

This raises, however, a lacuna in the Strawsonian approach when it comes to considering instances of evil understood in its narrow sense. It is

not clear that Shoemaker's tripartite account has the resources to understand the deficits of evil persons, such as sociopaths. It is striking that commentators in this area – from Strawson onwards – sometimes conflate those who we would regard as morally innocent, such as young children, from those who we would regard as morally deficient, such as sociopaths. In outlining Type-2 Pleas, for instance, Strawson suggests that they consist of excusing factors such as 'he was warped or deranged, neurotic or just a child'. Likewise, Gary Watson states that a 'child can be malicious, a psychotic can be hostile, a sociopath indifferent'. It seems discordant somehow to equate the 'warped or deranged' behaviour of a psychotic adult, or the indifference of a sociopath, with the bad behaviour of a child.

Extreme moral evil will likely have distinct roots from moral immaturity. The immature agent is excused – to the extent that they are excused – because they are not capable of clear deliberation; the child, we tend to think, should not be held fully responsible for their actions and attitude. There is a capability deficit in the child which is a mitigating factor. This is distinct from the agent who engages in extreme moral evil. Although some have tried to argue this way, it would be galling, at least on a standard intuition, to regard the sociopath's moral deficit – their incapacity to have any sympathetic regard for others – as a mitigating factor; it seems, rather, to be an aggravating factor. The fact that they cannot care about other's interests seems to inflame rather than defuse the moral reactions of others. As American philosopher Gary Watson observes, when discussing the actions of the serial killer Robert Harris, 'Harris's form of evil consists in part in being beyond the boundaries of the moral community'. Yet being beyond the boundaries of the moral community – in the sense of not yet being mature enough to properly participate in it – seems to be exactly what excuses or mitigates the child's behaviour. What explains these seemingly opposed reactions? Why should the child's incapacity mitigate, and the sociopath's incapacity aggravate? Are we simply, as Watson suggests, caught in an insoluble conflict where 'we are unable to command an overall view of [the evildoer's] life that permits the reactive attitudes to be sustained without ambivalence'?

An alternative suggestion to Watson's ambivalence about responsibility, and Shoemaker's pluralism about responsibility, would be to regard responsibility as partially contextual. On this view, rather than regarding

the internal criteria for being held responsible (in any of the three variations mentioned by Shoemaker) as statically attached to a particular strength of quality of will – such that, for example, one must have this much capacity for emotional sympathy to be eligible to be held accountable for any of their actions – we should, rather, regard distinct situations as calling for distinct strengths of quality of will.

For instance, suppose that one is faced with a very sensitive, vulnerable person who needs help. To help them one needs to have an exceptional capacity for sympathy. One needs to have a very strong quality of will of regard for other people. A person without such a strong quality of will of regard – a person without that exceptional capacity for sympathy – would not be blameworthy if they were not able to help, and perhaps even harmed through lack of needed regard, the vulnerable person. On the other hand, a person with a strong quality of will of regard – a strong capacity for sympathy – would be blameworthy (perhaps in the sense of being held accountable) if they were not to help, or, through lack of needed regard, were to harm that person. In some ways this is similar to how we think about child's responsibilities: their lack of capacity for standard adult-like regard for others is what mitigates them.

Conversely, some acts – or omissions – require a very limited capacity for sympathy. It takes very little sympathy or quality of will of regard to recognise that, for example, torturing an innocent person is harmful and should not be carried out. Most people who might be outside of the usual run of 'strains of involvement' of interpersonal interactions are not so far outside the moral community that they cannot understand or recognise the harm in such acts. The sociopath, who might be excused minor misdemeanours of regard on account of an unchosen deficit of affective sympathy, is nevertheless blameworthy if they carry out an extreme act such as torture. They are blameworthy because such acts require only a very weak quality of will of regard to realise that they are beyond the boundaries of acceptable conduct, and such a weak quality of will can be considered to be within the gift of such agents.

On the view painted above, then, the internal criteria of our participatory moral practice incorporate more than one tier. In the first tier, in the ordinary run of things, one expects and demands of others a particular strength of quality of will (which might be dispersed across

three varieties as mentioned by Shoemaker). An agent without such a quality of will – or with only a weakened version – might be excused from blame and other participant reactive attitudes relating to that deficit. While the agent with such a quality of will, but who does not exercise it, is blameworthy for their attitude and actions. In the second tier, in the extraordinary run of things – at the extremes of moral action (typically sadism, murder, cruelty) – the criteria for involvement is much lower. An agent with even a weakened quality of will can still be held accountable, attributable, and answerable for their actions, even though – exactly because of that weakened quality of will – they might be excused along one or more of the three dimensions for comparatively more minor, but typically blameworthy, acts. We expect all agents, even those with serious deficits and some diminished responsibility within ordinary interactions, to have the capacities to avoid doing evil – that is part of the criteria for involvement even with marginal cases – and those who fail to live up to those expectations can rightly be blamed for them, and their actions deemed evil.

The above provides an account of how agents can be held morally responsible for narrowly evil acts within a Strawsonian Framework that responds to the objections that typically arise with scepticism about the narrow concept of evil. The Strawsonian Framework neither requires recourse to a metaphysically suspect conception of evil (the Argument from Supernaturalism), nor leads to a process of dehumanisation (the Argument from Dehumanisation). The moral responsibility of an extreme evil doer can be explained by appeal to acts that would be blameworthy even for people who might be excused blame for less extreme acts. In this respect, agents who would usually seem as if they are outside the scope of the moral community for ordinary purposes, are drawn (back) into it when the acts are extreme enough.

What the account has not yet done is provide a definition of evil, nor try to determine what, within the Strawsonian framework, would serve to distinguish evil acts from what is merely wrong or bad. It would take us too far afield to discuss that in any detail here, so I shall make do with a very brief summary. The broad claim is that actions are considered evil to the extent that they are a continuation or strengthening along some dimension – typically the harm of another – of what is typically considered

bad and wrong, with the added element that they have a *distorting effect* on the moral framework itself. As the South African political theorist Stephen de Wijze has put it, evil acts or moralities 'invert or annihilate the "moral landscape"… needed for any civilised attempt to manage conflict and to establish a minimal level of respect and dignity between persons'.

The distorting effect, in my view, is that the nature of the wrong – the extremity of the act – is such that it prevents the kind of normative corrective that is typical of interactions in the wake of ordinary wrongdoing. With ordinary wrongdoing, there are criteria, internal to the practice, for what American philosopher Margaret Urban Walker calls *moral repair*: blame in the face of wrongdoing, remorse in the face of blame, forgiveness in the face of remorse, and reconciliation in the face of forgiveness. These provide the agents involved with a normative map for how they can find their way past a wrongdoing to a form of reconciliation. With extreme wrongdoing – that is, with evil – this normative corrective of moral repair is lost and the path to reconciliation becomes clogged, perhaps permanently, with no normative criteria provided for how to unclog it. In the face of evil, the scope for genuine remorse and genuine forgiveness is limited, and there are no normative criteria or expectations in place for how to deal with the act. As we have already observed, the evildoer who commits the act is liable to already be outside the scope of ordinary practices of interaction, including those of moral repair: we do not expect the genuinely morally disturbed to offer themselves up for genuine remorse. If there is to be remorse it is not to be expected or sought from, say, the cold-eyed sociopath. Likewise, if there is to be forgiveness – from the victims or their representative – then it cannot be normatively demanded: Forgiveness in the face of extreme wrongdoing can only be elective, and not conditioned by the norms of a practice. After all, it seems wrong to suppose that the victim of an atrocious crime might be required to forgive, even in the face of genuine remorse.

Evil and AI

The Strawsonian Framework provides an understanding of our moral practices in terms of a participatory practice involving reactive attitudes that respond to the attitudes and actions of others. This basic framework

– of ordinary interactions and ordinary moral repair – can then be expanded to include the following exceptions or amendments:

Exemption Cases: For some agents it is necessary to step-back and exempt them entirely from the 'strains of involvement' of the participatory stance. In such cases, where one takes the 'objective stance' towards another, that agent is not treated as a functioning member of the moral community but is treated as something to be 'managed'; a matter of 'policy' rather than of interactive regard.

Marginal Cases: For some agents it is necessary to treat them as responsible agents and members of the moral community in some respects but not in other respects. They might have, for example, the capacity for empathetic understanding, but not the capacity for careful deliberation; this will make a difference to the respect in which they are and are not responsible for their actions.

Extreme Cases: For some agents it is necessary to treat them as weakly responsible agents and as members only of the moral community only in respect of a certain sub-set of extreme acts. Such agents might, for example, have a highly reduced capacity for affective empathy which exempts them from some forms of responsibility in ordinary interactions, and outside of the ordinary forms of moral community, but it does not exempt them from those forms of responsibility when the acts are extreme.

We can now turn to the question of how this framework applies to AI and the respects in which AI might be considered morally responsible, including possessing a capacity for extreme evil.

Let's recall the three types of agency, outlined earlier:

(a) AI as a *non-agential tool*
(b) AI as a *hybrid, minimal agent*
(c) AI as a *fully autonomous agent*

We can see here, using the Strawsonian Framework, that when an AI is properly regarded as a non-agential tool – either temporally or permanently – then this is an *exemption case*. An exemption case will occur when there is no capacity, or use, in engaging with the machine as an agent with whom one might have a participatory, moral relationship. Encounters with such a machine involve taking the 'objective stance' which 'promotes the purely objective view of the agent as one posing problems simply of intellectual understanding, management, treatment and control'. The machine might be highly sophisticated and capable of operating independently from a human user. But, if it lacks the capacity to properly engage within interpersonal relations in a way that is at least minimally responsive to the morally infused attitudes of others, then one's attitude towards that machine can be one of management, control and repair. Note that this is the way that many debates about AI are framed: AI is not treated as a potential moral partner, but as a potential problem to be controlled; hence the ubiquitous term 'the control problem'.

When an AI is regarded as a hybrid, minimal agent, then questions of responsibility might go in one of two directions. The minimal AI agent could be understood as being strongly morally responsible along some dimensions, but not at all (or only marginally so) along others. The system might, for instance, be capable of calculating the means for a task, but require that its ends (or goals) be programmed into it. This would be a marginal case. For instance, if the AI in question is a semi-automated self-driving car, then it might be held *answerable* for its choices – why did it choose to career into the sidewalk when faced with oncoming traffic, rather than slamming on the brakes? –but needn't be *accountable* since the actions it performed were not an element of its 'deep self'.

What is interesting here, though, for our concerns, is where an AI hybrid, minimal agent is only weakly responsible along one or more dimensions. For instance, suppose that the AI has a strong capacity for deliberation about means – allowing for engagement within practices relating to answerability – but has only a weak capacity for affective empathy. This latter quality moves the AI into the territory of *a moral 'uncanny valley'*. We can see here a respect in which a mixed and weakened capacity for human-like responses to others; a mixed and weakened form of quality of regard and quality of will, would involve an agent who

engages in forms of ordinary participation in interpersonal relations – at least among some dimensions – but yet might, on account of a weakened capacity for affective empathy, or a weakened capacity for a strong deep self, be much more capable of causing extreme harm to others. Such an agent could be causally responsible for such acts on account of their having weak affective empathy, and also morally responsible for such acts on account of such acts requiring only weak affective empathy in order for the agent to be fully morally responsible for them: in other words, such a hybrid agent would be capable of extreme moral evil.

It is worth comparing that with the third form of agency: *a fully autonomous agent*. Although it is the AI as fully autonomous agent that has generated the most excitement and the most fear, it is my contention othat, in fact, it is the hybrid agent which should be considered the most liable to be capable of extreme moral evil. The reason for this is that a fully autonomous AI agent would be – on account of its being fully autonomous and responsible along the three dimensions – much more integrated within our everyday, interactive practices. If the Strawsonian Framework outlined here is correct, then the more that an AI system is integrated within our practices, then the more that system will be (correctly) understood as being responsible and autonomous in the relevant respects. The danger is not with such integrated, autonomous agents – it is with the semi-autonomous, hybrid agents: agents who have enough of a capacity to operate within a minimal form of human practices of participation and responsibility, yet remain partially outside of it, capable of acting from outside the parameters of ordinary interactions in certain circumstances, perhaps sometimes for reasons unknown. Here we have the capacity for genuine, extreme moral evil.

The concept of evil often conjures up images of the uncanny, of the macabre, of intentions that go beyond what Hannah Arendt called 'humanly comprehensible motive[s]'. It doesn't just arouse feelings of moral horror, but also arouses feelings of confusion and a sense of incomprehension. Although I do not believe that such uncanniness is definitional of evil – Arendt's 'banality of evil' thesis perhaps tells us that much – it is very often an accompaniment to it. The reason that such uncanniness accompanies evil, according to the ideas sketched here, is that it allows such actors to have one foot in our moral practices, while having

one foot outside of them. The concern with AI is that, in our current trajectory, we seem to be creating systems with that kind of hybrid capacity, of being responsible in some ways, and less so in other ways. A way to remedy this would be to try to fully integrate artificial systems into our forms of participatory practices – ensure that such agents are able to engage in practices of reactive attitudes with the capacity for moral repair – rather than partially integrating them, and then hoping for the best.