P&D Philosophy & Digitality
An Open Access, Peer-Reviewed Journal

From knowing too little to knowing too much

automated emotion recognition and implications for autonomous agency

Lydia Farina

Abstract: If one accepts that affect has an important role to play in reasoning and deliberation, the ability to accurately recognise and predict emotions and associated behaviour by Artificial Intelligence Systems (AIs) allows for cases of manipulation and restriction of agency. To support this claim, I give a brief discussion on cases of 'knowing too little' where AIs are tasked with making predictions on the agents' emotions without having enough information to guarantee that these predictions would be accurate. As with other issues in AI research, such issues may dissolve with the advance of technological progress. In the main part of the chapter, I deal with issues arising from AIs being able to accurately recognise human emotion and predict actions associated with those emotions. I call these cases of 'knowing too much' to highlight issues relating to manipulation. I show that in these cases, there is a restriction of agency such that it becomes difficult to claim that the agent is autonomous or that they can be held responsible for any resulting actions or behaviour. I conclude that solving the problem of 'knowing too little' will inadvertently lead to the creation of the second problem and suggest some important questions for further research.

Keywords: emotion; autonomous agency; affective computing; emotion recognition; manipulation

1. Introduction

1 One of the claims made by Picard to justify the interest in affective computing was the importance of emotion in reasoning, deliberating and decision making (Picard 1997, 2015). In the last 30 years or so the role of emotion in reasoning has been under considerable debate (Damasio 1994; De Sousa 1987; Frank 1988; Goldie 2004; Brady 2013; Tappolet 2016). In this paper I argue that, if one accepts that: 1) emotion has an important role to play in reasoning and deliberation and 2) we design Artificial Intelligence Systems (AISs) with access to agents' emotions and behaviour (in terms of recognising and predicting chains of emotions and associated behaviour), this ultimately allows cases of agent manipulation. In turn, this manipulation has direct implications for autonomous agency.

2 AISs are applications or technologies using Artificial Intelligence tools and techniques to perform specific tasks e.g. to find patterns in data, make predictions or decisions, recognise images etc. (Peters 2022). To give some specific examples, recommender systems determine communications to users on social media feeds or provide scores for interviewes during interviews, algorithms operate fitness wearable devices and conversational AIs are used in customer services (Peters 2022; Klenk 2021; Bogen 2019). This paper focuses on possible cases where AISs could use emotion recognition techniques to make predictions on agents' emotions and associated behaviour with the aim to encourage or discourage certain behaviours. This influence on an agent's emotions and behaviour admits degrees from cases of minimum influence to cases of manipulation. It would be good to note here that the manipulation claim does not imply that there is no possible positive outcome from AISs having the ability to influence emotion. Instead, I accept that such a capability can have positive applications and implications: for example, in health environments where the individual for whatever reason cannot control the intensity of an emotion, it may be beneficial to have an application which can warn the human agent or nudge them away from certain emotion/behaviour chains to help the human agent control or regulate this emotion.

Contact: Lydia Farina <lydia.farina@nottingham.ac.uk>
https://doi.org/10.18716/ojs/pd.v2i2.11948

In these situations, the individual would prefer to be nudged away or in any case be more able to control the intensity of those emotions when it has been associated with inappropriate or life endangering behaviour.

3 In this paper, I focus on cases of manipulation rather than cases of pure influence. In everyday life situations, we influence each other by acting one way or another or recommending actions to other agents without this influence being considered outright morally bad. For example, even though you do not normally wake up early to go for a run, you have told me in the past that you think it would be beneficial for you to run regularly and that you do not like to run on your own. One morning I send you a text to invite you to join me for a morning run and you accept. My action is influencing your behaviour and I am letting you know that I am doing this to help you achieve a goal. Or alternatively you ask me to let you know when I do go for a morning run each time I go so that you are motivated to go for a run. It is plausible to think in this and similar cases that the agent is influenced but not manipulated. On the other hand, in cases of manipulation, an agent is manipulated when the link between her mental states and associated behaviour is interfered with in such a way that she suffers an injustice (Klenk 2021). Borrowing from Klenk's account, I take manipulation to occur in cases where an injustice causes an agent to have mental states which would not have existed but for the injustice occuring. To give an example used typically in the literature on manipulation, in Shakespeare's *Othello,* the main character is manipulated by Iago's lies to feel jealous and in an enraged state, kills his wife Deisdaimona. Othello is manipulated by Iago in the sense that his mental states of anger, jealousy or rage are caused by and are based on falsehoods; injustice occurs because he is obstructed from knowing the real reasons behind Iago's words and actions.

4 Although an agents' behaviour is frequently influenced by their environment and other agents, what separates influence from manipulation on this account is that in manipulation cases, agents do not have access to the reasons behind the influencer's behaviour. This may be because the influencer refuses to give them or conceals them, as in Iago' s case. However, we can also think of cases where these reasons are not visible because of negligence or a lack of care associated with whoever is causing the injustice (Klenk 2021). For example, one may be spreading misinformation on social media which influences others because they have a complete lack of care about whether this information is true or not. This is different to cases of disinformation where an agent knowingly spreads lies on social media (Nguyen 2020) to cause specific behaviours or beliefs to others. However, in both cases an injustice is caused because the manipulated agents have certain mental states e.g. desires, beliefs, emotions which they would not have had, had manipulation not occured. Although manipulation can occur through influencing both cognitive and affective mental states, in this paper I focus on influencing affective mental states and more specifically manipulating affective mental states to encourage or discourage behaviour.

5 If one wishes to manipulate someone's emotions, it would be helpful if one had the ability to recognise these emotions both before and after any attempt to manipulate to confirm the attempt was successful. In everyday life interactions, we typically recognise other's emotions through information provided from facial expressions, an agent's behaviour or their testimony. In lab environments or when using technological devices, one can also have access to information relating to physiological responses e.g. skin conductivity or brain activation information through FMRi, EEG etc. (Fragopanagos and Taylor 2005). These types of information can be used for recognising human emotions as in identifying what emotional state a human agent is experiencing. There is considerable debate on the accuracy of emotion recognition by both humans and AISs and associated challenges such as cultural variability and context.[1] If accurate emotion recognition is an important condition for the ability to manipulate affective states efficiently, in cases where this is missing, so in cases of 'knowing too little', it is not guaranteed that manipulation will occur in an efficient way e.g. in the sense that the manipulator recognises that manipulation has or has not occurred.

6 Largely, cases of 'knowing too little' can result from the complex nature of emotion, a degree of unpredictability associated with human behaviour, and the lack of current technological capability to correctly recognise or classify human emotion in the wild and so outside laboratory and experimental spaces.[2] However, cases of 'knowing

1. See for example Waelen (2024); Cabitza et al. (2022); Vaccaro et al. (2020); Barrett et al. (2019); Kragel et al. (2018); Kreibig & Gross (2017); Jack (2016); Hess & Hareli (2016); Scarantino et al. (2015); Gendron et al. (2014); Hassin et al. (2013); Lindquist et al. (2012); Aviezer et al. (2008).

2. See Cabitza et al. (2022) for a review of the current accuracy of emotion recognition techniques used by AISs.

too little' may dissolve with the advance of technological progress.[3] For example, emotion recognition accuracy can improve by using personalisation techniques (Nwadike et al. 2024; Pearson 2019). These techniques aim to improve emotion recognition by training AISs to recognise the facial expression of a single individual; the hope is that by learning the individual's typical facial expressions rather than her exaggerated facial expressions, this may help increase accuracy scores. As such, it is possible that by using improved techniques we will be able, at some point in the future, to accurately recognise human emotion and predict the behaviour arising from or associated with them.

7 In the main part of the chapter, I accept this possibility and deal with issues arising from future AISs being able to accurately recognise human emotion and robustly predict actions associated with those emotions. I call these cases of 'knowing too much' to highlight issues relating to manipulation. I show that in these cases there is a restriction of agency such that it becomes difficult to claim that the agent is autonomous. The degree of manipulation can be exacerbated because 1) human agents are not aware of what types of AISs are used and to what purposes and 2) the belief that manipulation can only occur when the influencer intends to manipulate. As argued by Klenk (2021), the latter is not necessary because manipulation can also occur because of lack of care or negligence. I conclude that solving the first problem of 'knowing too little' may inadvertently lead to the creation of the second problem of 'knowing too much' and suggest some important questions for further research.[4]

8 I include below my claim as a set of premises which, if likely to be true, entail the conclusion:
P1. AISs (properly designed) can manipulate the emotions and behaviour of human agents including inducing or regulating emotion.
P2. Emotion influences practical deliberation in human agents viewed as the ability for self-reflective evaluation.
P3. From 1 and 2 it follows that 'AISs can manipulate (restrict or interfere with) practical deliberation of human agents' (by manipulating their emotions and associated behaviour).
P4. Practical deliberation provides justification for attributions of autonomous agency.

———————

Conclusion: Artificial Intelligence Systems (properly designed) can restrict autonomous agency by reducing or removing the capacity to provide justification for actions.

9 In the following sections I provide evidence to support premises 1 and 2 included above by discussing the role of emotion in self-reflective evalution and autonomous agency and by looking at possible cases where knowing too much about an agent's emotions and being able to manipulate these states, restricts an agent's autonomous agency.

2. Autonomous Agency and Self-Reflective Evaluation

10 Personal autonomy is identified as the capacity to evaluate one's motives, values and goals by considering one's beliefs and desires (Christman 1991; Mele 1993). If this evaluation shows that the motives do not align with one's beliefs and desires, the motives need to be adjusted accordingly. An agent is a being with the capacity to act intentionally such that their actions are causally related to their intentions (Davidson 1980; Bratman 1987; Dretske 1988; Mele 2003).[5] Combining these definitions, an autonomous agent is a being with the capacity to act intentionally and a capacity to evaluate the motives of intentional action to consider whether they align with held beliefs and desires.[6]

3. For current studies showing progress in this area, see Kuipers et al. (2023); Nicolai et al. (2022); Saffaryazdi et al. (2022); Küntzler et al. (2021); Loaiza (2021); Storbeck et al. (2015). For older studies showing that some emotions are recognised with enough frequency in certain contexts, see Ekman (1972); Ekman & Friesen (1971); Ekman & Cordaro (2011).

4. Even if one is pessimistic about technological progress in the future, if our current technological tools can recognise human emotion in some cases and can create associations between emotions and an agent's behaviour, this still allows for cases of manipulation to be possible. This means that the main claim of this paper also applies to cases where AISs have these capabilities on the basis of information available to them currently.

5. For alternative accounts of autonomous agency, see Ginet (1990) or Lowe (2008) who do not define agency in terms of causal relations between agent-involving states and events.

6. This account of autonomous agency allows cases where the agent can still act autonomously even when they hold false beliefs about the world.

11 The capacity to evaluate one's motives in light of held beliefs and desires is also described as a capacity for self-reflective evaluation. Through self-reflection agents can check whether an action aligns with their motives and if not, correct any misalignment cases. If someone or something interferes with the agent's ability for self-reflective evaluation, for example because of conditioning or indoctrination, this removes or cancels out the agent's autonomy (Buss and Westlund 2018). The capacity for evaluation presupposes a capacity for self-reflection; an agent must reflect on their own values, goals, desires and beliefs when deliberating how to act. Self-reflective evaluation takes place when one deliberates about how to act; it provides evidence for autonomous agency and the exercise of freedom of the will (Frankfurt 1971; Bilgrami 2006). If one finds a way to cancel or interfere with this capacity, one can restrict autonomous agency.

12 Importantly, and as discussed above, a big part of one's ability for evaluative self-reflection during deliberation is based on their ability to have good knowledge of their self in the sense of knowing their emotional traits and dispositions as well as knowledge of their values, goals etc. This knowledge is important during deliberation because it helps agents determine which action aligns with their values and goals and which does not. This knowledge also includes knowing what actions they are usually motivated towards when undergoing an emotion and what actions/behaviours they are motivated to avoid. To give an example, I may avoid spending time with an individual who expresses sexist or racist ideas because I do not wish to experience the emotions associated with this interaction e.g. anger, frustration etc. or the behaviours I portray in these encounters e.g. having to explain that racism is bad etc.

13 The importance of emotion for self-reflective evaluation is also evident in the relation between emotion and personal identity or character (Goldie 2000). An important part of what makes us who we are, is the ability to control or regulate our emotions but also the ability to choose what emotions we allow ourselves to engage with, in what manner and to what degree. Losing control over our emotions and the behaviours associated with them means that we lose control over our personal identity. Having control over our emotions according to these views gives us control of the self-reflective evaluation process and of our personal identity. The relevance of emotion here is that according to views focusing on personality or moral motivation (e.g. Goldie (2000); Tappolet (2016)), emotions are a fundamental feature of what makes a person the person that she is; her emotional repertoire along with her attitudes towards it, are a big factor in how she behaves and what kind of personality she displays.[7] To show how emotion can influence an agent's actions in the example mentioned above, if I know that a certain individual will make me feel angry because of racist comments, I can avoid meeting that individual because I do not want to experience the emotions usually elicited by this encounter and the actions I choose to perform e.g. disagree, get into a fight etc. Generally speaking, the specific environments we find ourselves in are very important for the emotional repertoire we are encouraged to develop and maintain in the sense that they either encourage or discourage specific emotional behaviours (Colombetti 2017). Where we have control over the environments we find ourselves in, this facilitates the control over our emotions. However, where we have minimal or no control over the environments we find ourselves in, this removes our ability to control our emotions.

14 One can claim that regardless of the environment, if the agent is not forced to do an action suggested by someone else, this allows for some degree of autonomous agency. According to this objection, if there is a degree of autonomy still available, this ensures that they are autonomous agents and that they can be held responsible for their actions. This objection goes against a wealth of research in philosophy and psychology showing that the environment one is living in shapes one's character and behaviour. For example, situationist studies show that in cases where we do actually act freely, the degree of freedom is significantly reduced by the situation and environmental context we find ourselves in (Darley and Latane 1968; Milgram 1963; Haney et al. 1973). In such cases, primes in the environment may induce certain behaviour or decision without conscious acknowledgement by the agent that this is taking place (Wegner and Wheatley 1999; Wegner 2002). In the later cases, as shown in the experiments by Wegner & Wheatley (1999), the agent may report that they consciously willed a certain behaviour/action when we know that they did not. This makes it possible that manipulation can occur even when the agents are not explicitly forced to choose a specific behaviour.

7. If an individual's identity 'derives from the role she plays as the central character in the story of her life' (Kind (2015, page 125); Schechtman (1996)), then not playing that central character would mean that the agent loses control of their personal identity.

2.1 Emotion and Self-Evaluative Reflection

15 There are different ways one can influence an agent's ability for self-reflective evaluation. One way is to stop the process of evaluation outright for example by persuading the agent that such a process is not necessary so that they act without exercising this capacity. An alternative way would be to change the process of evaluation itself such that it still goes on but not as a guiding principle towards the decision; instead, its function is to provide justification for the decision after the decision has been reached. To explain how this can take place, I borrow from accounts of moral deliberation which take emotions to be essential in this process. According to these accounts, when we deliberate about what the moral thing to do is, what we are actually doing is trying to find justification for a moral judgement existing already. This moral judgment is identical to how we feel, and so to our emotions about what we are deliberating. Moral deliberation in this sense is a process we undertake when we are trying to rationalise judgments we have reached already on the basis of our emotions. According to this view, emotions are not complimentary to the process of moral deliberation; instead they are actually constitutive of or identical with moral judgements (Haidt 2001; Richardson 2018). When we are deliberating about what to do, emotions and reason are not co-creating this decision; instead, how we feel about things *just is* how we select to act. Reason provides a justification for the selected action as a *post hoc* exercise engaged to justify a decision we have already made rather than an exercise to help us reach this decision. If these strong views about the role of emotion in moral deliberation are correct and emotions are indeed constitutive of and identified with judgements, the ability to induce or encourage specific emotions to agents can lead to controlling the process of moral deliberation itself. If one controls the process by which agents decide how to act, they restrict autonomous agency.[8]

16 Alternative accounts of moral deliberation accept that emotion plays an important part in this process without also accepting the stronger view that emotions are identical with moral judgments. According to these accounts, emotions help us during the evaluation process by making our values more salient; they are perceptions of our values (Tappolet 2016). In this respect, they help motivate a specific action rather than another and so ultimately help us align actions with our values and beliefs. Because of the direct link between our emotions and our values or desires, we are in a better position to know which actions are compatible with them (Tappolet 2016). Another way to describe a person's values, motives and goals is to talk about someone's traits and dispositions, what is sometimes called one's character (Goldie 2000). In order for an agent to be in a position to self-reflectively evaluate their actions, they need to have knowledge of their own character which includes values, goals, motives, traits and dispositions. In this way knowledge of ones' own character along with the capacity for self-reflection are essential features of human agency (Taylor 1985).

17 If one accepts this latter account, one way to interfere with the process of self-reflective evaluation is to control or influence the desires, values or goals the agent accepts as their own. In this case, the agent would still be allowed to reflect on whether the motives align with her desires, but she does not freely choose these desires as her own. Here the discussion moves from the agent being free to evaluate whether her motives align with her desires during deliberation to whether the agent freely chooses which desires, goals and values she will consider her own. Typically, the latter is considered as a fundamental feature of someone's personal identity in the sense that, in theory, we can select which goals, desires or values we wish to associate with our personality or character. However, and as argued by Meyers and others, societal influences can interfere in such a way that some agents cannot choose freely which goals and desires they can adopt in forming their own personal identity (Meyers 2002; Neisser and Jopling 1997).

18 Regardless of which view on the role of emotion one accepts from the views discussed above, the role of emotion during self-reflective evaluation is important: a) if one accepts the strong account, manipulating one's emotions provides a direct route for manipulating their moral judgements. b) if one accepts the alternative account, manipulating one's emotions has a direct effect on the adoption of goals, values, motives and ultimately the agent's character. To conclude, if one accepts that emotion plays an important role for the capacity

8. Although the examples used in this paper focus on moral reasoning, because of the motivational force of emotion any implications may be applicable to reasoning in general.

of agents to evaluate actions using self-reflection, manipulating emotion allows a direct route to cancelling or controlling the process of self-reflective evaluation during delibration.

3. Manipulation of Emotions by AISs

19 Technologies oriented towards emotion recognition can use different types of information from textual analysis, clicking, velocity of scrolling, eye tracking to information on facial expressions of emotion by using Facial Expression Recognition Systems (Khare et al. 2024; Samadiani et al. 2019). For example, FERS are used in healthcare to help diagnose anxiety (McClure et al. 2003), or in virtual and augmented reality technologies to enhance the quality of the interaction (Hickson et al. 2017; Chen et al. 2015). For a more recent example, FERS are used in research projects producing Adaptive Interactive Movies (AIM) such as the film 'Before we Disappear' (2024) where machine learning technologies coupled with computer vision applications interpret each spectator's moods and emotions to adapt the film in real time and choose one of possible three endings.[9]

20 In this section, I look at cases where through successful emotion recognition and personalisation AISs could manipulate emotions and associated behaviours in agents. Agreeing with Baumann & Döring (2011), I argue that emotion oriented technologies could restrict autonomous agency. This restriction becomes problematic in cases where 1) the user would not consent to it if consent was required, 2) the user is completely unaware that it takes place or 3) the user may be aware that automated emotion recognition is taking place but is unaware of associated operations e.g. predictions of behaviour and the reasons or purposes of those operations. In the latter cases, the owner of the application or the owner of the platform the application is being used on, may state that they are using this technology to provide more tailored experiences to users by recognising their preferences through emotion recognition or to monitor performance or wellbeing in the workplace (Suni Lopez et al. 2019; Gal et al. 2020; Spataro 2020). To show why these cases may restrict autonomous agency and the capacity for self-reflection, I discuss a hypothetical case below. Although this is a hypothetical scenario, I include references showing cases where AISs are already used to perform some of the activities mentioned. This moves the scenario from the realm of fictional cases to the realm of plausible contemporary cases.

3.1 Hypothetical case

21 We can think of a possible scenario where an AIS, for example a recommender feeder on a digital platform, is tasked with eliciting as many behaviours/actions of type x in the human agent (HA) without having acquired relevant consent by the HA or without revealing the purposes behind using this technology.[10] The AIS is not designed with any ethical or moral code which would stop it from using any available means to generate specific types of behaviour. In this scenario, and as is typically the case in digital platforms, the end goal of the recommender system is to elicite specific behaviours by controlling both the content of the feed and the time allocated to each feed with the ultimate goal to increase commercial profit.[11] There are several ethical risks associated with using recommender systems such as risks to privacy and opacity. In this paper, I focus on the risks associated with autonomous agency and personal identity.[12] It is reasonable to accept that the AIS has the ability to scaffold the human agent's environment e.g. by recommending specific events, facts, opinion articles, comments, videos, images etc. and attract the individual's attention to specific content. This scaffolding creates specific affordances for the user in the sense that the content chosen by the recommender system encourages the elicitation of specific emotions to the user.[13] In general, affordances are interpreted as possible actions enabled by the environment (Gibson 1979). More specifically affordances in social media technologies (SMTs) are 'the relational properties of SMTs that are likely to induce an emotional state or emotion-related behavior' (Steinert and Dennis 2022). If the AIS through using automated recognition can identify that specific contents elicit

9. This AIM was co-produced by Dr Ramchurn and Blueskeye AI in 2024, see Icke (2023).

10. For a comprehensive review of recommender systems and the ethical risks associated with them see Milano et al. (2020).

11. For a discussion of the different types of recommender systems and the tasks they usually perform see Burr et al. (2018).

12. Milano et al. (2020) suggest a taxonomy for risks associated with recommender systems. The risk to autonomous agency and personal identity discussed in this paper falls under the 'rights violations' category in their taxonomy.

13. This is similar to the impact that comments or information from users of social media platforms can have on the social platform users. As Steinert & Dennis (2022, page 36) describe, social media technologies 'facilitate online emotions because of emotional affordances'.

emotions in agents and others do not, they may use the former to capture the agent's attention. Automated emotion recognition in this case would be based on processing information on sentiment analysis referring to text, clicking, velocity of scrolling or FERSs.[14] The ability to attract the agent's attention is an important step in this process. If the AIS can also be trained via personalisation to make successful associations between identified emotional states and correlating behaviour, they can then predict an agent's behaviour which then gives them a direct route to bringing this behaviour about.

22 If the above scenario is possible, the AIS is in a position to attract the HA's attention and elicite a specific emotion to the HA. Which emotion the AIS will select, depends on the behaviour associated with that emotion. For example, if the HA spends more time on a specific digital platform when feeling fear or anger, the AIS could select to elicite fear or anger in the HA via some communication so that the HA spends more time on the digital platform. Alternatively, the AIS may identify that the HA is currently undergoing a specific emotion, e.g. anger and then send them photos or articles or comments from specific platforms to enhance this anger and promote engagement with these platforms such as reading comments by others, making comments etc.

23 In the scenario mentioned above, AISs could induce specific emotions in agents where they have been trained to know that these emotions are associated with specific behaviours in these agents and they are tasked with bringing about a specific behaviour. In other words, if they know that this behaviour robustly follows a specific emotional state, and they also know what cues usually bring this emotional state about, they can use these cues to bring the emotional state about. To give an example, if the AIS is tasked with increasing agent's A screen time on a particular digital platform and they are in a position to know which emotional states increase screen time on the platform and also which cues bring about these emotional states in agent A, they can increase A's screen time by recommending content which elicits these emotional states. The capability for manipulation can increase through personalisation techniques when AISs become very efficient in recognising what emotions a specific agent is experiencing and can associate these emotions with behaviours e.g. with experiencing other emotions, spending time on specific websites etc.

3.2 Manipulation of autonomous agency

24 The case above would qualify as a case of manipulation because the agent is not aware of the reasons why the recommender system is recommending this content to them. Injustice in this case can arise because of lack of consent, lack of knowledge of the mechanisms underlying this interaction, or lack of declaring the commercial profit motivating the manipulation.

25 The risk to autonomous agency in these cases is associated with the ability of the AIS not only to predict certain behaviours but also to cause them. For example, the AIS can induce feelings of anger or frustration by recommending an article on a specific issue which has caused you feelings of anger on a previous occasion. In that occasion, you posted an angry comment on social media which caused controversy and then resulted to your comment getting many responses. With this knowledge, the AIS can decide that it is advantageous for the task it is currently performing e.g. to increase your screen time on the platform, to induce the same type of behaviour. To give other examples, the AIS can recommend an article asking you to sign a political petition out of a newly induced moral judgment about the content of it, or to vote towards a certain goal, or to sign up to a particular group advocating the newly induced moral judgement etc. Knowing that you are prone to do these things whilst feeling angry, may give it the power to manipulate your behaviour one way or another. As such, it becomes pertinent to ask whether agents lose their autonomous agency, in part or wholly, when they are manipulated into specific behaviours by AISs.

26 Cases where AISs accurately recognise and predict an agent's behaviour show that they know 'too much' about an agent; in these cases, AISs can be used to restrict autonomous agency. If emotion plays an important part in self-reflective evaluation during deliberation either by interfering or by controlling it, the combination of 1) the ability to induce emotions to an agent and 2) knowledge of associations of emotion-action in an agent, can lead to manipulation and restriction of autonomous agency. If we design AISs with a capability to recognise

14. It currently is not likely that many users of online platforms would consent to giving access to facial expressions during use of platforms but automated emotion recognition can be successful by processing other types of available information as mentioned earlier.

and predict human emotion, perhaps by combining applications already operating in this area, this will enable their human owners to be in a position of power compared to the human agents these AISs are developed to interact with. In those cases, manipulation takes place when the AIS either induces or enhances an emotion and in this way is able to increase the probability of specific motivations and actions associated with this specific emotion.[15] This ability goes in both directions e.g. it is an ability to bring about certain actions, behaviours or choices, but also it is an ability to stop or nudge away from certain emotions and actions, behaviours of choices arising from these emotions.

27 If the scenario above is plausible, by improving currently existing emotion recognition techniques and combining them with other techniques such as personalisation, we allow manipulation cases in environments where AISs interact with human agents. Agreeing with Milano et al. (2020), such manipulation cases can be considered as rights violations because they restrict or abuse autonomous agency. Further research in this area would help determine rights and responsibilities for all stakeholders in these environments along with necessary steps to avoid violation of rights and harmful interactions.

4. Conclusion

28 Affective states are considered as either influencing or constituting moral judgments and their importance for moral reasoning and self-reflecctive evaluation is widely accepted in the literature (Haidt 2001; Tappolet 2016; Richardson 2018). If one recognises an agent's emotional dispositions and knows which action typically follows an emotion and vice versa, one can restrict autonomous agency by interfering with the process of moral deliberation and self-reflective evaluation. If one designs and uses AISs to recognise and induce emotions and associated behaviour to human agents, this has direct implications for the agent's capacity for self-reflecive evaluation during moral deliberation and ultimately for autonomous agency.[16] Further research is needed in this area to determine the checks and balances that need to be put in place to avoid cases of manipulation and other harmful types of interaction.

Note

29 I would like to thank the anonymous reviewers of this issue for the constructive comments and the attendees of ISRE2024 Conference at Queen's University Belfast for the helpful suggestions on an earlier version of this paper.

References

Aviezer, Hillel, Ran R. Hassin, Jennifer Ryan, et al. 2008. "Angry, Disgusted, or Afraid?: Studies on the Malleability of Emotion Perception." *Psychological Science* 19 (7): 724–32. https://doi.org/10.1111/j.1467-9280.2008.02148.x.

Barrett, Lisa Feldman, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements." *Psychological Science in the Public Interest* 20 (1): 1–68. https://doi.org/10.1177/1529100619832930.

Baumann, Holger, and Sabine Döring. 2011. "Emotion-Oriented Systems and the Autonomy of Persons." In *Emotion-Oriented Systems*, edited by Roddy Cowie, Catherine Pelachaud, and Paolo Petta. Cognitive Technologies. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-15184-2_40.

15. I accept that the manipulation claim does not need to apply to all cases and that there will be some human agents who will not be manipulated. In addition, some agents may be manipulated in some instances but not in others. However, the claim is not that this manipulation takes place in all cases. Instead, the claim is that this manipulation is possible in some cases for some agents.

16. This can be used as another restrictive tool which exacerbates existing oppression and power inequalities.

Bilgrami, Akeel. 2006. *Self-Knowledge and Resentment*. Harvard University Press.

Bogen, Miranda. 2019. "All the Ways Hiring Algorithms Can Introduce Bias." Havard Buisness Review, May 16. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias.

Brady, Michael S. 2013. *Emotional Insight: The Epistemic Role of Emotional Experience*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199685523.001.0001.

Bratman, Michael E. 1987. *Intention, Plans, and Practical Reason.* Harvard University Press.

Burr, Christopher, Nello Cristianini, and James Ladyman. 2018. "An Analysis of the Interaction Between Intelligent Software Agents and Human Users." *Minds and Machines* 28 (4): 735–74. https://doi.org/10.1007/s11023-018-9479-0.

Buss, Sarah, and Andrea Westlund. 2018. "Personal Autonomy." In *The Stanford Encyclopedia of Philosophy*, Spring 2018, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/.

Cabitza, Federico, Andrea Campagner, and Martina Mattioli. 2022. "The Unbearable (Technical) Unreliability of Automated Facial Emotion Recognition." *Big Data & Society* 9 (2): 20539517221129549. https://doi.org/10.1177/20539517221129549.

Chen, Chien-Hsu, I.-Jui Lee, and Ling-Yi Lin. 2015. "Augmented Reality-Based Self-Facial Modeling to Promote the Emotional Expression and Social Skills of Adolescents with Autism Spectrum Disorders." *Research in Developmental Disabilities* 36 (January): 396–403. https://doi.org/10.1016/j.ridd.2014.10.015.

Christman, John. 1991. "Autonomy and Personal History." *Canadian Journal of Philosophy* 21 (1): 1–24. https://doi.org/10.1080/00455091.1991.10717234.

Colombetti, Giovanna. 2017. "Enactive Affectivity, Extended." *Topoi* 36 (3): 445–55. https://doi.org/10.1007/s11245-015-9335-2.

Damasio, Antonio R. 1994. *Descartes' Error: Emotion, Reason and the Human Brain*. G.P. Putnam's Sons.

Darley, John M., and Bibb Latane. 1968. "Bystander Intervention in Emergencies: Diffusion of Responsibility." *Journal of Personality and Social Psychology* 8 (4, Pt.1): 377–83. https://doi.org/10.1037/h0025589.

Davidson, Donald. 1980. "Agency." In *Essays on Actions and Events*, by Donald Davidson. Clarendon Press. https://doi.org/10.1093/0199246270.003.0003.

De Sousa, Ronald. 1987. *The Rationality of Emotion*. 1. paperback ed., 5. print. Bradford Books. MIT Press.

Dretske, Fred. 1988. *Explaining Behavior: Reasons in a World of Causes*. The MIT Press. https://doi.org/10.7551/mitpress/2927.001.0001.

Ekman, Paul. 1972. "Universals and Cultural Differences in Facial Expressions of Emotion." In *Nebraska Symposium on Motivation*, edited by James K. Cole, vol. 19. University of Nebraska Press.

Ekman, Paul, and Daniel Cordaro. 2011. "What Is Meant by Calling Emotions Basic." *Emotion Review* 3 (4): 364–70. https://doi.org/10.1177/1754073911410740.

Ekman, Paul, and Wallace V. Friesen. 1971. "Constants across Cultures in the Face and Emotion." *Journal of Personality and Social Psychology* 17 (2): 124–29. https://doi.org/10.1037/h0030377.

Fragopanagos, N., and J. G. Taylor. 2005. "Emotion Recognition in Human–Computer Interaction." *Neural Networks* 18 (4): 389–405. https://doi.org/10.1016/j.neunet.2005.03.006.

Frank, Robert H. 1988. *Passions Within Reason: The Strategic Role of Emotions*. 1. ed. Norton.

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68 (1): 5–20. https://doi.org/10.2307/2024717.

Gal, Uri, Tina Blegind Jensen, and Mari-Klara Stein. 2020. "Breaking the Vicious Cycle of Algorithmic Management: A Virtue Ethics Approach to People Analytics." *Information and Organization* 30 (2): 100301. https://doi.org/10.1016/j.infoandorg.2020.100301.

Gendron, Maria, Debi Roberson, Jacoba Marietta Van Der Vyver, and Lisa Feldman Barrett. 2014. "Perceptions of Emotion from Facial Expressions Are Not Culturally Universal: Evidence from a Remote Culture." *Emotion* 14 (2): 251–62. https://doi.org/10.1037/a0036052.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Ginet, Carl. 1990. *On Action*. 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9781139173780.

Goldie, Peter. 2000. *The Emotions: A Philosophical Exploration*. Clarendon Press.

Goldie, Peter. 2004. "Emotion, Reason and Virtue." In *Emotion, Evolution, and Rationality*, edited by Dylan Evans and Pierre Cruse. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198528975.003.0013.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–34. https://doi.org/10.1037/0033-295X.108.4.814.

Haney, Craig, Curtis Banks, and Phil Zimbardo. 1973. "A Study of Prisoners and Guards in a Simulated Prison." *Naval Research Review* 9: 1–17.

Hassin, Ran R., Hillel Aviezer, and Shlomo Bentin. 2013. "Inherently Ambiguous: Facial Expressions of Emotions, in Context." *Emotion Review* 5 (1): 60–65. https://doi.org/10.1177/1754073912451331.

Hess, Ursula, and Shlomo Hareli. 2016. "The Impact of Context on the Perception of Emotions." In *The Expression of Emotion*, 1st ed., edited by Catharine Abell and Joel Smith. Cambridge University Press. https://doi.org/10.1017/CBO9781316275672.010.

Hickson, Steven, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. 2017. "Eyemotion: Classifying Facial Expressions in VR Using Eye-Tracking Cameras." arXiv:1707.07204. arXiv, July 28. https://doi.org/10.48550/arXiv.1707.07204.

Icke, Jane. 2023. "Films That Watch You – Interactive Technology Turns the Viewing Experience on Its Head." University of Nottingham, February 23. https://www.nottingham.ac.uk/news/films-that-watch-you-interactive-technology-turns-the-viewing-experience-on-its-head.

Jack, Rachael E. 2016. "Cultural Specificities in the Transmission and Decoding of Facial Expressions of Emotion." In *The Expression of Emotion*, 1st ed., edited by Catharine Abell and Joel Smith. Cambridge University Press. https://doi.org/10.1017/CBO9781316275672.009.

Khare, Smith K., Victoria Blanes-Vidal, Esmaeil S. Nadimi, and U. Rajendra Acharya. 2024. "Emotion Recognition and Artificial Intelligence: A Systematic Review (2014–2023) and Research Recommendations." *Information Fusion* 102 (February). https://doi.org/10.1016/j.inffus.2023.102019.

Kind, Amy. 2015. *Persons and Personal Identity*. Key Concepts in Philosophy 1. Polity.

Klenk, Michael. 2021. "Manipulation, Injustice, and Technology." The Philosophy of Online Manipulation 3883189. Edited by Fleur Jongepier and Michael Klenk. Social Science Research Network, October 14. https://doi.org/10.2139/ssrn.3883189.

Kragel, Philip A., Leonie Koban, Lisa Feldman Barrett, and Tor D. Wager. 2018. "Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging." *Neuron* 99 (2): 257–73. https://doi.org/10.1016/j.neuron.2018.06.009.

Kreibig, Sylvia D., and James J. Gross. 2017. "Understanding Mixed Emotions: Paradigms and Measures." *Current Opinion in Behavioral Sciences* 15 (June): 62–71. https://doi.org/10.1016/j.cobeha.2017.05.016.

Kuipers, Mithras, Mitchel Kappen, and Marnix Naber. 2023. "How Nervous Am I? How Computer Vision Succeeds and Humans Fail in Interpreting State Anxiety from Dynamic Facial Behaviour." *Cognition and Emotion* 37 (6): 1105–15. https://doi.org/10.1080/02699931.2023.2229545.

Küntzler, Theresa, T. Tim A. Höfling, and Georg W. Alpers. 2021. "Automatic Facial Expression Recognition in Standardized and Non-Standardized Emotional Expressions." *Frontiers in Psychology* 12 (May): 627561. https://doi.org/10.3389/fpsyg.2021.627561.

Lindquist, Kristen A., Tor D. Wager, Hedy Kober, Eliza Bliss-Moreau, and Lisa Feldman Barrett. 2012. "The Brain Basis of Emotion: A Meta-Analytic Review." *Behavioral and Brain Sciences* 35 (3): 121–43. https://doi.org/10.1017/S0140525X11000446.

Loaiza, Juan R. 2021. "Emotions and the Problem of Variability." *Review of Philosophy and Psychology* 12 (2): 329–51. https://doi.org/10.1007/s13164-020-00492-8.

Lowe, E. J. 2008. *Personal Agency: The Metaphysics of Mind and Action*. 1st ed. Oxford University PressOxford. https://doi.org/10.1093/acprof:oso/9780199217144.001.0001.

McClure, Erin B., Kayla Pope, Andrea J. Hoberman, Daniel S. Pine, and Ellen Leibenluft. 2003. "Facial Expression Recognition in Adolescents With Mood and Anxiety Disorders." *American Journal of Psychiatry* 160 (6): 1172–74. https://doi.org/10.1176/appi.ajp.160.6.1172.

Mele, Alfred. 1993. "History and Personal Autonomy." *Canadian Journal of Philosophy* 23 (2): 271–80. https://doi.org/10.1080/00455091.1993.10717320.

Mele, Alfred. 2003. *Motivation and Agency*. 1st ed. Oxford University PressNew York. https://doi.org/10.1093/019515617X.001.0001.

Meyers, Diana Tietjens. 2002. *Gender in the Mirror: Cultural Imagery and Women's Agency*. 1st ed. Oxford University PressNew York. https://doi.org/10.1093/0195140419.001.0001.

Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. 2020. "Recommender Systems and Their Ethical Challenges." *AI & SOCIETY* 35 (4): 957–67. https://doi.org/10.1007/s00146-020-00950-y.

Milgram, Stanley. 1963. "Behavioral Study of Obedience." *The Journal of Abnormal and Social Psychology* 67 (4): 371–78. https://doi.org/10.1037/h0040525.

Neisser, Ulric, and David A. Jopling, eds. 1997. *The Conceptual Self in Context: Culture Experience Self Understanding*. 1. publ. Emory Symposia in Cognition 7. Cambridge Univ. Press.

Nguyen, C. Thi. 2020. "Echo Chambers and Epistemic Bubbles." *Episteme* 17 (2): 141–61. https://doi.org/10.1017/epi.2018.32.

Nicolai, Susanne, Philipp Franikowski, and Susanne Stoll-Kleemann. 2022. "Predicting Pro-Environmental Intention and Behavior Based on Justice Sensitivity, Moral Disengagement, and Moral Emotions – Results of Two Quota-Sampling Surveys." *Frontiers in Psychology* 13 (June): 914366. https://doi.org/10.3389/fpsyg.2022.914366.

Nwadike, Munachiso, Jialin Li, and Hanan Salam. 2024. "Improving Personalisation in Valence and Arousal Prediction Using Data Augmentation." arXiv:2404.09042. arXiv, April 13. https://doi.org/10.48550/arXiv.2404.09042.

Pearson, Andrew. 2019. "Personalisation the Artificial Intelligence Way." *Journal of Digital & Social Media Marketing* 7 (3): 245–69. https://doi.org/10.69554/JJGR7331.

Peters, Uwe. 2022. "Algorithmic Political Bias in Artificial Intelligence Systems." *Philosophy & Technology* 35 (2): 1–23. https://doi.org/10.1007/s13347-022-00512-8.

Picard, Rosalind W. 1997. *Affective Computing*. MIT Press.

Picard, Rosalind W. 2015. "The Promise of Affective Computing." In *The Oxford Handbook of Affective Computing*, edited by Rafael Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199942237.013.013.

Richardson, Henry S. 2018. "Moral Reasoning." In *The Stanford Encyclopedia of Philosophy*, Fall 2018 Edition, edited by Edward N. Zalta. August 27. https://plato.stanford.edu/archives/fall2018/entries/reasoning-moral/.

Saffaryazdi, Nastaran, Syed Talal Wasim, Kuldeep Dileep, et al. 2022. "Using Facial Micro-Expressions in Combination With EEG and Physiological Signals for Emotion Recognition." *Frontiers in Psychology* 13 (June): 864047. https://doi.org/10.3389/fpsyg.2022.864047.

Samadiani, Najmeh, Guangyan Huang, Borui Cai, et al. 2019. "A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data." *Sensors* 19 (8): 1863. https://doi.org/10.3390/s19081863.

Scarantino, Andrea, Lisa Feldman Barrett, and James A. Russell. 2015. "Basic Emotions, Psychological Construction, and the Problem of Variability." In *The Psychological Construction of Emotion*. The Guilford Press.

Schechtman, Marya. 1996. *The Constitution of Selves*. 1st ed. Cornell University Press.

Spataro, Jared. 2020. "The Future of Work—the Good, the Challenging & the Unknown." Microsoft 365 Blog, July 8. https://www.microsoft.com/en-us/microsoft-365/blog/2020/07/08/future-work-good-challenging-unknown/.

Steinert, Steffen, and Matthew James Dennis. 2022. "Emotions and Digital Well-Being: On Social Media's Emotional Affordances." *Philosophy & Technology* 35 (2): 36. https://doi.org/10.1007/s13347-022-00530-6.

Storbeck, Justin, Nicole A. Davidson, Chelsea F. Dahl, Sara Blass, and Edwin Yung. 2015. "Emotion, Working Memory Task Demands and Individual Differences Predict Behavior, Cognitive Effort and Negative Affect." *Cognition and Emotion* 29 (1): 95–117. https://doi.org/10.1080/02699931.2014.904222.

Suni Lopez, Franci, Nelly Condori-Fernandez, and Alejandro Catala. 2019. "Towards Real-Time Automatic Stress Detection for Office Workplaces." In *Information Management and Big Data*, edited by Juan Antonio Lossio-Ventura, Denisse Muñante, and Hugo Alatrista-Salas, vol. 898. Communications in Computer and Information Science. Springer International Publishing. https://doi.org/10.1007/978-3-030-11680-4_27.

Tappolet, Christine. 2016. *Emotions, Values, and Agency*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199696512.001.0001.

Taylor, Charles. 1985. "Self-Interpreting Animals." In *Human Agency and Language*, vol. 1. Philosophical Papers. Cambridge University Press. https://doi.org/10.1017/CBO9781139173483.

Vaccaro, Anthony G., Jonas T. Kaplan, and Antonio Damasio. 2020. "Bittersweet: The Neuroscience of Ambivalent Affect." *Perspectives on Psychological Science* 15 (5): 1187–99. https://doi.org/10.1177/1745691620927708.

Waelen, Rosalie. 2024. "Philosophical Lessons for Emotion Recognition Technology." *Minds and Machines* 34 (1): 3. https://doi.org/10.1007/s11023-024-09671-3.

Wegner, Daniel M. 2002. *The Illusion of Conscious Will*. MIT Press.

Wegner, Daniel M., and Thalia Wheatley. 1999. "Apparent Mental Causation: Sources of the Experience of Will." *American Psychologist* 54 (7): 480–92. https://doi.org/10.1037/0003-066X.54.7.480.