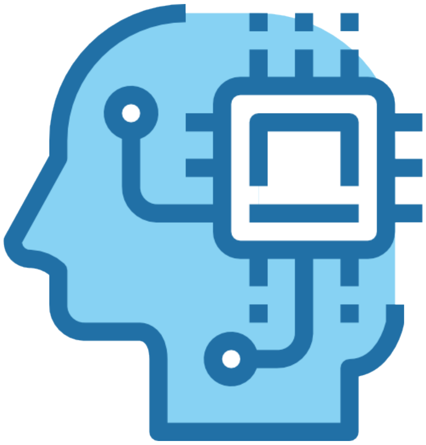

AI, Hume's Account of Human Action, and Moral Responsibility

Kekoa Wong
Computer Engineering and Philosophy
University of Notre Dame, Class of 2022

December 2021

Abstract

This paper explores the contradictions that arise in Hume's account of human action due to the moral responsibility of AI in today's society. Using his ideas of liberty, necessity, and moral sentimentalism, it will be argued that Hume's theory is a strong foundation for the science of human behavior, but is not sufficient in establishing appropriate extensions of moral responsibility due to the rise of autonomous agents.

1 Introduction

Hume develops a strong theory of human action based on his account of causality, drawing inferences about human decisions based on the nature of observable events in the world. Through this theory, Hume states that humans can only act with liberty when they are the source of their own action, using the determinations of their own will. Furthermore, he argues that humans can only hold sole moral responsibility when their decisions reflect such liberty. However, the rise of AI has produced challenges to Hume's concept of liberty and moral responsibility. Such agents can seemingly make autonomous decisions, yet the moral weight of their actions seems to fall on the application designers. An analysis of this extension of moral responsibility will be examined in this paper, along with the ramifications that it holds for Hume's theory of human action. First, Hume's theory will be reconstructed using his account of liberty and necessity while tying this theory to morality. Then, AI's capacity for liberty and responsibility will be analyzed in the context of Hume's argument. Next, this analysis will be used to illustrate the contradictions that arise from Hume's account. Finally, the paper will conclude that alternate theories to Hume's may be best suited for establishing moral responsibility in today's technological society, as they may be more effective at distinguishing between the action of artificial agents and humans.

2 Hume's Argument

Hume establishes a theory of human action that applies his account of causation and necessity to voluntary action. In the context of inanimate matter, Hume's account of necessity dictates that "matter, in all its operations, is actuated by a necessary force and that every natural effect is so precisely determined by the energy of its cause that no other effect, in such particular circumstances, could possibly have resulted from it"(Hume, Hume, 8.1.4). With this definition, all events in the natural world must have a necessary connection to a cause, being correlated in this way. Extending this account to human action, Hume argues for a form of psychological determinism that connects a person's motivations to their voluntary actions. He writes that "our actions have a constant union with our motives, tempers, and circumstances"(Hume, 1991, 2.3.1), situating these inputs as the cause that generates the effect of our action. These sentiments have a functional mapping with the output of human decisions, where "the same motives always produce the same actions" similar to how the "same events follow from the same causes"(Hume, Hume, 8.1.7).

However, Hume acknowledges that human sentiments are greatly diverse and open to change over time, writing that there may be a "gradual change of our sentiments and inclinations and the different maxims which prevail in the different ages of human creatures"(Hume, Hume, 8.1.11). But through this change, the sentiments and inclinations still contain a certain "uniformity in their influence," as we would never be able to gain a consistent observational understanding of the person if their conduct deviated immensely. Additionally, he grants that there may be the possibility that some actions may "seem to have no regular connection with any known motives and are exceptions to all the measures of conduct"(Hume, Hume, 8.1.12). These actions are unforeseeable events but by no means eliminate the possibility of psychological determinism. Instead, such actions are simply non deterministic since they are unobservable and have never been analyzed previously. However, once these actions exhibit themselves, a philosopher has an opportunity to discover the origins of this action and relate it back to the motivations that would have caused such an event, keeping inline with psychological determinism.

This account of human action causes implications for the definition of free will and liberty for intelligent beings. As a result of this theory, Hume can only define liberty as "a power of acting or not acting according to the determinations of the will"(Hume, Hume, 8.1.23). Thus, a decision is considered free only if it is the result of the agent's own motivations and not bound by the extrinsic motivations of an outside force. As Hume is a compatibilist, free will is not understood as the ability to make alternative choices with the same inputs, but instead only requires that the source of an action lies within the motivations and sentiments of the individual rather than the environment.

But where does this lead us toward decisions that entail moral responsibility? Hume argues that "actions are objects of our moral sentiment so far as they are indications of the internal character, passions, and affections"(Hume, Hume, 8.1.31). Furthermore, he claims that "it is impossible that they can give rise to either praise or blame where they do not proceed from these principles, but are derived altogether from external violence". In this way, humans must be expressing their liberty, acting in accordance with their own motivations, to carry moral responsibility for their actions. In other words, they can only be morally culpable when they are the source of the action, which reflects on the internal content of their causing motivations. However, when external pressures dictate the action, such as a prisoner being held in chains, the human is not culpable for their decision since it is derivable from the environment.

3 Inquiry into AI's Liberty and Moral Responsibility

But does this mean that any being of deterministic nature can hold moral responsibility on the account of their actions? For instance, a system of artificial intelligence would be deterministic in nature, being built from neural networks and a loss function. The agent would base its decision making on the data that it has gained from its environment, being constrained by its innate nature tied to the structure of its networks and the loss function that it is trying to optimize. In this way, AI can be compared to the intelligent being of a human, where its innate nature could be compared to the psychological predispositions of a human and the input data being related to the experience of the human that is the foundation of their decision making. Hume's idea of necessity is important here, as there is a necessary connection between the action and the motivating cause. In the context of AI, the cause of a decision must be necessarily connected to the algorithms and experiential data input of the AI itself. Therefore, AI could be considered an "intelligent agent," as it can facilitate decisions, expressing its liberty since it is the source of the action.

Using Hume's idea of morality tied to liberty, it could be inferred that AI should hold moral responsibility for its actions when making decisions that express its liberty in this way. The functional action of the system can be necessarily connected to the network structure, optimization pattern, or data, thereby illustrating that an AI is the source of its decision. With this necessary connection and liberty, why should it not be held solely accountable for the implications of its decisions? This inference would elicit deep ramifications across today's and tomorrow's society. AI algorithms continue to have a broad impact on decisions across many aspects of life, automating many without the input of humans. Therefore, how would one even go about holding responsibility for the implications of these actions on an artificial being? One can imagine the circumstance of an automated car deciding to run over many pedestrians to save the driver's life or killing the driver instead, exhibiting a modern trolley dilemma. Additionally, one can also conceptualize an automated algorithm recommending misinformation or inappropriate/violent content, harmfully impacting individuals and society in general. Due to the difficulties of holding an artificial being responsible, the weight of such actions has typically fallen on the shoulders of the human designers. But how could a Humean argue for this type of a response?

4 AI's Derived Passions from a Designer

Returning to Hume's idea of moral sentiments, one may argue that an artificial being is unfit to hold moral responsibility since it lacks the appropriate predispositions to lead to action and the composition of its internal characteristics are fundamentally different from humans. In A Treatise of Human Nature, Hume analyzes the interaction between passion and reason, arguing that "reason alone can never be a motive to any action of the will" and that "it can never oppose passion in the direction of the will" (Hume, 1991, 2.3.2). Hume proposes that reason can be compared to the mechanics that regulate "the motions of bodies to some designed end or purpose" (Hume, 1991, 2.3.2). With this argument, Hume establishes reason as the method that leads us towards a desired goal or trajectory that is set by our passions. However, reason in of itself can never be the pure motivator of our actions. In the same way, it can never prevent us from doing an action or act in opposition to a passion or impulse (Hume, 1991, 2.3.4).

Therefore, all human behavior must be motivated by a passion, whether that be toward action or inaction, including decisions of moral weight. Reason only provides the methodology to regulate between these passions in the consideration of a path. The power of AI lies in its ability to establish strong statistical reasoning based on the data that it has processed. Its internal characteristics are structured around providing powerful inferences for decisions based on input data. But the end motivation for this reasoning ultimately lies within the loss function parameters which it is built to optimize, and it lacks any true passions that would drive it toward actions (in the way humans do).

Thus, one may argue that an artificial system is not truly practicing liberty in its decision-making as it is enslaved to the loss function that a human designer establishes with their own sentiments. The sentiments and passions in an artificial system are missing, and it is simply building the reasoning, using statistics and neural networks, for the motivations of the designers who created the loss function. In this context, AI can only be perceived as a simple tool of reasoning that fulfills the duties of its architect. With this argument, moral responsibility is extended to the creator, as the origin of the sentiments lie with the motivations of this individual instead of the system of reasoning that they created.

5 Resulting Contradictions in Hume's Theory

But this line of argument creates a serious ramification to Hume's idea of liberty. It was established that an intelligent system like AI does not carry moral responsibility since the motivations of the system were not its own, being created by its designers. Yet, to what extent are human motivations, psychological predispositions, and passions their own and not a product of their creator? Due to Hume's idea of necessity, the cause of actions must have an origin. But what if there are certain humans who contain a psychological predisposition for violence, feeling high amounts of pleasure for such actions? Or consider the plight of individuals who suffer from anxiety, depression, or sociopathy, feeling irregular amounts of emotions in relation to other humans. Are these individuals really the source of such sentiments, or could these dispositions be compared to AI's enslavement to a loss function?

These humans carry sentiments that are a result of their innate bodily system and they are not responsible for the creation of such constraints. They may be forced to live with this reality, realizing that the cause of their actions can be traced back to these motivations. Therefore, does this mean that the human has liberty to express their true nature when their motivations are so tied up with these predispositions that they have no control over? Are they truly that different from AI? Similar to AI's loss function, one may argue that these dispositions resemble the motivations of the designer rather than the product itself. Therefore, the psychological predispositions expressed in the nature of humans must resemble the values of the creator since the human is not the necessary source of such motivations.

This line of argument creates significant consequences for the perfection of a god that is an essential characteristic for many Enlightenment thinkers. Hume realizes the ramifications that this argument creates for his religious world. He writes that "the ultimate Author of all volitions is the Creator of the world, who first bestowed motion on this immense machine and placed all beings in that particular position" (Hume, Hume, 8.2.32), acknowledging God as the first necessary source. He continues by addressing this counterargument, saying that "human actions, therefore, either can have no moral turpitude at all, as proceeding from so good a cause or, if they have turpitude, they must involve our Creator in the same guilt, while he is acknowledged to be their ultimate cause and author"(Hume, Hume, 8.2.32). Therefore, God must also be incriminated in any action holding moral value, and held responsible as the designer of a system that elicited such an action. Yet Hume believes that such a conclusion is an absurd consequence, as he believes that this Deity is perfect in these ways (Hume, Hume, 8.2.32). But since he still seeks to hold humans criminally responsible for their actions, he faces extreme difficulties in dispelling the notion of an imperfect God. Thus, he acquiesces to this contradiction in the face of an imperfect God, writing that "to reconcile the indifference and contingency of human actions with prescience or to defend absolute decrees and yet free the Deity from being the author of sin has been found up to now to exceed all the power of philosophy" (Hume, Hume, 8.2.36).

Such a conclusion admits the impossibility in reconciling the notion between innate predispositions and holding human beings solely accountable for their actions. Any being that contains such predispositions, and is unable to change them in a way that is isolated from the causes of their environment or creator, must share responsibility for their actions with a designer.

6 Argument for Alternative Theories

Therefore, Hume is faced with a contradiction in his account of action that arises when addressing the moral responsibility of AI. Simply put, if an agent has built-in dispositions that it is unable to change, then it must share moral responsibility with its designer. This line of reasoning is more palatable for a Humean argument when examining the case of AI and the ingrained sentiments of its creator. However, when expanding this argument toward humans and God, Hume is unwilling to open up the possibility for human indeterministic behavior (allowing them the sole casual power to change their own predispositions) or the idea of an imperfect God. Thus, while Hume's theory is a useful foundation to the science of human behavior, the implications of his account create contradictions that he is unable to alleviate.

In order to more effectively understand the type of action required for moral responsibility, we must turn to alternate theories. Some approaches that are open to the conception of an imperfect god would have potential, yet they would prevent humans from holding sole responsibility over their actions due to the existence of innate deterministic predispositions. This would mean that the agent's designer would always hold responsibility, allowing both AI and humans to escape full blame. Instead, exploring theories that allow humans the indeterminate power to alter their predispositions may be more promising. In this way, human action would be distinctly different from artificial action in that it would separate them from a deterministic being such as AI. Humans would have the power to overcome their innate predispositions under these theories, whereas AI would continue to be constrained to its ingrained loss function. Therefore, AI's designer would hold moral responsibility over its actions while humans would hold responsibility over their own actions. This approach would enable us to understand the decision making from different types of beings and would more adequately encapsulate the moral responsibility that can be held by artificial agents and humans.

7 Conclusion

Overall, Hume's theory is useful in the scientific study of human behavior, allowing us to seek to understand the complex motivations behind our actions. Yet, there are significant contradictions that arise due to the presence of innate predispositions, the extension of moral responsibility, and his idea of a perfect God. These contradictions are especially apparent in the context of today's autonomous agents and the extension of moral responsibility for their actions. Thus, to understand a being's capacity for moral responsibility, theories of action that allow for indeterminate behavior must be considered.

References

Hume, D. An Enquiry Concerning Human Understanding. In *Modern Philosophy: An Anthology of Primary Sources*. Indianapolis: Hackett Publishing Company, Inc.

Hume, D. (1991). *A treatise of human nature*. Oxford: Clarendon Pr.