Hume's Guillotine in Designing Ethically Intelligent Technologies

Pertti Saariluoma(✉)

Cognitive Science, University of Jyväskylä,
Box 35, 40014 Jyväskylä, Finland
ps@jyu.fi

Abstract. Intelligent machines can follow ethical rules in their behaviour. However, it is less clear whether intelligent systems can also create new ethical principles. The former position can be called weak ethical AI and the latter strong ethical AI. Hume's guillotine which claims that one cannot derive values from facts appears to be a fundamental obstacle to strong ethical AI. The analysis of human ethical information processes provides clarity to the possibility of strong ethical AI. Human ethical information processing begins with positive of negative emotions associated to situations. Situations can be seen as consequences of actions and for this reason people can define rules about acceptability of typical actions. Finally, socio-ethical discourse create general ethical rules. Intelligent systems can provide important support in ethical process and thus the difference between weak and strong ethical AI is polar.

Keywords: Interaction design · Intelligent systems · Ethical design · Hume's guillotine

1 Introduction

Once again, humankind is on the cusp of a new technology revolution. It has encountered such situations many times in its history. Technologies, work processes, and societies have changed numerous times. Stone tools, fire, sailing, navigation, cannons, printing, clocks, steam engines, electricity, and nuclear energy provide good examples of technological revolutions leading to new forms of work and social organisation [1]. The ongoing revolution is based on intelligent technologies. They are characterised by their capacity to carry out tasks, which have previously required the intelligent information processing of the human mind.

The improved speed of computing and the fast growth of data have made it possible to design technical artefacts with the capacity to do tasks, which thus far only people have been able to carry out. Modern examples of emerging intelligent technologies are not few. Artificial intelligence has penetrated numerous aspects of modern life. Industrial robots, office automation, intelligent medicine, changes in teaching, autonomous traffic systems, and intelligent finance give us a fast vision of the future [2, 3].

In addition to fast routine processing of logical inferences, machines can make decisions between alternative courses of action. They can even learn to make classifications of their own so that people are not able to predict the information states which

T. Ahram et al. (Eds.): IHIET 2020, AISC 1152, pp. 10–15, 2020.
https://doi.org/10.1007/978-3-030-44267-5_2

intelligent systems can generate. Consequently, intelligent systems can select between different sense-making courses of actions.

The capacity for selective information processing makes it possible for modern AI-based systems to compare the values of different information states on sense-making grounds. A chess-playing computer for example, can find the best sequences of moves among millions of legal alternatives. Intelligent choices make a machine's actions intelligent.

A very specific view is opened by ethics with respect to thinking intelligent choices. Some information states are more ethical than others, and thus it makes sense to discuss ethics in the context of acting intelligent machines. They can select some courses of action as they are more ethical with respect to certain ethical principles. Thus, intelligent technologies can make operational decisions on ethical grounds. They can choose between different courses of actions on the grounds of implemented ethical principles. For example, intelligent systems can prefer children to middle-aged people in making decisions about the order of medical operations. Such decisions are ethical and carried out by intelligent machines.

For the reasons given, one can speak of ethics typical to using intelligent technologies in two senses. One can speak of the ethical use of technical artefacts in society, but one can also develop systems with ethical capacities of some type. In this paper, focus is in the latter.

2 Hume's Guillotine

A crucial question in considering future intelligent sociotechnical society is how machines can be ethical at all. They are just systems with different electric states which people map as information about reality. The electric states are mapped to factual information. In digital systems power is either higher or lower, and this makes it possible to have two states. These states can stand for truth or false. Thus, information in intelligent machines is apparently factual. Intelligent machines process facts.

Facts are different from values. While facts are binary and can be true or false, values are not dichotomous. Something can be obliged, forbidden, or allowed. The problem of relations to binary facts in binary machines and multiple state values is important in designing ethical information systems and is conceptually important in designing ethically intelligent technologies.

One important problem in relations to facts and values was seen over 250 years ago. Hume [4] wrote: "It is impossible that the distinction between moral good and evil can be made by reason". This aporia is called Hume's guillotine or "is-ought to" problem, which is central to modern ethics. Hume's guillotine claims that one cannot derive from how things are how they should be. When designing ethically intelligent machines, Hume's guillotine is a relevant conceptual problem. One can justly ask whether machines processing facts can have anything to do with ethics at all, and if they do, how is it possible?

Intelligent machines can be ethical in more than one sense. The first position is that people implement their values in the evaluative structures of ethical programmes as traditional chess machines have their human implemented heuristics. This latter

position can respectively be termed *weak ethical AI (WEAI).* The position that machines are able to generate new ethical rules and principles themselves can be called *strong ethical AI or intelligence (SEAI).* In the context of the former position, Hume's guillotine is apparently easier to solve than in the latter. However, firstly, it is important to ask how ethical information processing is possible for people and how weak and strong ethical AI differ from each other.

3 Ethical Process

Hume's guillotine is still an important ethical dilemma today, and one cannot say that it has been solved. To gain clarity on this issue, one must think how it is possible for people to process ethical information. The idea that human information processing can be used to develop intelligent technologies has been called cognitive mimetic [5]. Here, the analysis of human ethical information processing can be used as a model for respective machine information processing.

Ethics are possible as they are real. There are no grounds to doubt that people are capable of creating ethical rules and norms. The process of creating ethical rules and norms can be called an *ethical process* or *ethical information process*, which is an example of human creative thinking. Ethical machines are machines which can participate in an ethical process.

Human experience, i.e. conscious mental representation, forms a central component of human information processing and thinking. The information contents of experiences and representations can be called mental contents. Mental representations have their cognitive and emotional dimensions. Both have an important role in ethical information processing, but Hume's guillotine cuts them apart.

Ethically, an important type of mental content is emotional valence [6]. Most emotions can be divided into positive or negative, pleasant or unpleasant, and happy or sad. Therefore, all situations emerging in the course of actions can be experienced positively or negatively.

Emotionally grounded ethical thinking is normally labelled as emotivism [7]. These theories begin with the idea that situations of life and respective experiences are emotionally positive or negative (pleasant and unpleasant). The emotional analysis of consequences of actions thus provides the basis for the ethical analysis of actions and action types. For example, the so-called golden rule (one should not treat others in ways that one would not like to be treated oneself) can be seen as a generalisation of situational experiences of deeds in which the principle is followed or violated. Thus, the emotional and ethical information process is in the analysis and experience of the emotional valence and can be taken as the first point of the ethical process.

Consequently, the development of ethical norms is grounded in the analysis of emotional situations. However, it is not wise to end the analysis of the ethical process with emotions. The situations of life are consequences of actions. Thus, the value of actions can be defined on the grounds of the valence of the situations arising as a consequence of particular types of actions. Norms describe what kinds of actions have had emotionally positive or emotionally negative consequences. Actions leading to pain are not acceptable and actions leading to positive emotions are good.

The first step in defining ethical principles is to classify situations emotionally and actions leading to situations of two types respectively good or bad. Thus, one can generate ethical norm "avoid excessive use of alcohol as it leads to social and health problems". Alcoholism is a situation in life and drinking is the action which ends to this situation.

However, different people experience situations in different ways. Social interaction can be painful for some while it is positive for another. Therefore, the general ethical norms can be seen to be consequence of informal (everyday) and formal (or political) discourses. This socio-ethical process has been investigated in discourse ethics % [8]. Thus, it is essential to add to the ethical process the discourse between people in society related to political analysis and even laws.

Hume missed that the ethical process and each norm in the generation process has three components. Firstly, there is an emotional analysis of situations in life. People do this kind of analysis every moment of their life. Secondly, the ethical process includes factual analysis of actions leading to the given types of situations. Finally, one needs to add a socio-ethical discourse, which defines the social and historical properties of a situation. Though Hume understood clearly the triad of emotions, reason, and action, his guillotine unreasonably broke the process.

Hume's guillotine is a consequence of a mistaken analysis of the ethical process and ethicality of actions. Hume does not pay attention to the fact that ethics arise from the simultaneous analysis of situations. Cognitive and emotional aspects of situations are encoded in a parallel manner. This is why, the very question whether (cognitive) facts be used to define (emotional) values is senseless. Facts and values are two sides of one and the same mental event. Social discourse works to get a generalised idea about the relations of actions, cognitions, and emotions. Accurate analysis of the ethical process makes it possible to study the problems of weak and strong AI from a new perspective.

4 Weak and Strong Ethical AI

The analysis of the ethical process aids us in considering the relations of weak and strong ethical AI. Following the founding ideas of life-based design giving clarity to the way ethics and ethical norms are created in human life enables researchers to study the generation's ethical design requirements and the ethical information processing for technologies. Searching for answers to two questions is central. Firstly what kinds of technologies should be developed, and secondly, how can these technologies be taken part of everyday life [9].

Weak AI is not a difficult case. Ethical norms can be implemented in AI programmes. It is possible to define the situation and their factual properties. This information can be recognised by intelligent systems in data, and associate ethical norms can be followed in actions. Thus, designers can build recognition association type action models with ethical contents. For example, if some situation is known to cause pain, technology should act to avoid such situations.

However, strong AI is more challenging, and there are no clear-cut solutions to the problems of designing strong ethical AI. Actually, the border between weak and strong

AI is not absolute, but systems can differ in their strength. The criterion for the strength of an ethical AI (EAI) system is the capacity to create new ethical norms without human involvement. Firstly, it is possible by means of data analysis to study possible pain or negative valence causing situations. For example, data mining can find factors causing illnesses, which have been unknown so far. Such research has existed for a long time. For example, Durkheim [10] found a link between religions, social discourse, and suicides, and a connection between smoking and lung cancer was found in the sixties. There is no logical obstacle to finding such associations by means of intelligent systems. Thus, human-supported AI and data mining can be used to find novel factual grounds for new ways of behaving. This kind of EAI is machine-supported AI.

Another possibility is to ask machines to recognise features, which are known to cause emotionally negative experiences. It is also possible to register human responses to different types of situations to classify them as emotionally negative. AI programmes can actively search for new combinations so that the human component is one-step further from the previous case. The information found can be associated with the actions ending in negative situations, and thus new information can be used to create new ethical rules.

Finally, the core issue is whether intelligent systems can create new previously unknown ethical norms without human involvement to process on the grounds of their factual data. Machines can analyse by different means emotional valences typical to some situations. They can also associate the results of emotional analysis to the actions. They can even analyse general social attitudes in these situations. The autonomy of ethical systems can thus be gradually increased. But human involvement can be relatively direct in creating new ethical rules.

5 Final Discussion

Since information systems are involved in carrying out increasingly complicated actions. It is essential to develop ethical capacities for these systems. Their operational roles can be very independent, and thus it is essential that they can follow sense-making ethical practices.

Apparently, Hume's guillotine can make it hard to develop ethical autonomy for future systems. Intelligent systems are in the first place factual information processing devices, and it is not easy to see how one could derive values from facts. Despite conceptual difficulties, it is important to think how intelligent systems can follow ethical norms in their actions.

Our analysis suggests that there seems to be two poles in ethical information processing, which can be called weak and strong ethical AI. The first kind of system can apply given ethical rules in given situations. They can recognise critical features in situations and choose their actions on the ground. In such cases, ethics are just a human implanted feature in a recognition action system. This kind of ethical processor can be called weak AI.

Nevertheless, despite Hume's guillotine, people are able to create ethical thoughts and information processes. Thus, it must be possible to create machine-supported

ethical processes with greater autonomy. The analysis of the ethical process also provides clues about how machines can be added to improve and create existing ethical processes. Thus, the second pole in the strength of ethical AI is formed by systems which can collect data, associate it with situations, and link the situations to emotional valence and respective actions. Finally, such systems could develop new ethical principles to follow. This kind of ethical AI can be called strong. Thus, developing strong ethical AI is a gradual process, and there are no absolute limits between its weaker and stronger forms.

References

1. Bernal, J.: Science in History. Penguin Books, Harmondworth (1969)
2. Ford, M.: Rise of the Robots. Basic Books, New York (2015)
3. Tegmark, M.: Life 3.0. Penguin Books, Harmondsworth (2017)
4. Hume, D.A.: A Treatise of Human Nature. Dent, London (1972/ orig. 1738)
5. Saariluoma, P., Kujala, T., Karvonen, A., Ahonen, M.: Cognitive mimetics: main ideas. In: Arabnia, H.R., Fuente, D.D.L., Kozerenko, E.B., Olivas, J. A., Tinetti, F. G. (eds.), ICAI 2018: Proceedings of the 2018 International Conference on Artificial Intelligence, pp. 202–206. CSREA Press (2018)
6. Frijda, N.: Emotions. Cambridge University Press, Cambridge (1986)
7. Malik, K.: A Quest for a Moral Compass. Atlantic Books, London (2014)
8. Habermas, J.: Diskursethik (Discourse Ethics). Surkamp, Frankfurth am Main (2018)
9. Saariluoma, P., Cañas, J., Leikas, J.: Designing for Life. Macmillan, London (2016)
10. Durkheim, E.: Suicide. Free Press, New York (1951)