# Hume's Guillotine Resolved

Pertti Saariluoma

Jyväskylä University, 40014 Jyväskylä, Finland
ps@jyu.fi

**Abstract.** According to Hume's guillotine, one cannot derive values from facts. Since intelligent systems are fact processors, one can ask how ethical machines can be possible. However, ethics is a real-life process. People analyze actions and situations emotionally and cognitively. Thus they learn rules, such as "this situation feels good/bad." The cognitive analysis of actions is associated with emotional analysis. The association of action, emotion and cognition can be termed a primary ethical schema. Through an ethical information process in which emotions and cognitions interact in social discourse, primary ethical schemas are refined into ethical norms. Each component of the process is different, but they cooperate to construct an ethical approach to thinking.

Hume's guillotine mistakenly breaks down primary ethical schemas and juxtaposes emotions and cognitions. There is no ethics without coordinated emotional, cognitive and social analysis. Therefore, his theory can be seen as a pseudo problem.

In the future, ethical processes will involve intelligent systems that can make ethical choices. Weak ethical artificial intelligence (AI) systems can apply given ethical rules to data, while strong ethical AI systems can derive their own rules from data and knowledge about human emotions. Resolving Hume's guillotine introduces new ways to develop stronger forms of ethical AI.

**Keywords:** Hume's guillotine, intelligent system, ethical information processes

## 1 Introduction

Human living centers on satisfying basic human needs. People set goals and use technical artefacts (e.g. programs, machines and devices) to help them pursue these goals. Thus technical artefacts, and how people use them, are important to consider when constructing the future intelligent information society. Designers must fit new technologies into the basic contexts of human life.

People act in groups and unify their efforts to reach their goals socially. They constantly participate in an infinite number of explicit and implicit action systems – what I call "forms of life" – such as programmers, soccer fans, nurses, workers, newspaper readers, elevens and executives. The concept of a form of life allows us to apply a I recommend clarifying what you mean by this.

single concept to discuss kindergartens as well as parliaments and religious cultures [1] [2]. Given the structure of human actions, it is helpful to consider human technology

interaction (HTI) problems from the human point of view. One of the central issues related to HTI today is ethics for intelligent systems.

Ethics can be seen as a system of ethical or moral rules and principles or laws [3]. However, the concept can also be considered as a system of real human actions [4]. For example, one can use ethical rules to ethically regulate marriage (i.e. normative ethics or norm-oriented ethics), but it is also possible to investigate how people in different cultures *really* act in their marriages (i.e. real-life or life-based ethics) [4].

When ethically regulating the human use of (or interaction with) intelligent systems, both axiological and action-based thinking are relevant. However, if researchers take the latter path, it will be essential to understand ethics as human information processing, i.e., thinking about what happens in human minds when they interact with intelligent systems among other people.

Human actions are controlled by mental processes: the mind tells the body what to do. Experience provides people with conscious representations of situations, actions and feelings; underlying these experiences, the mind operates as an information processing system in a broad sense [5] [6]. People experience *cognition*, which refers to information-related processes such as perceiving, attending, remembering and using language and thinking, and *emotions*, which include feeling and appraising the goodness and badness of things and events in relation to oneself [7]; researchers and designers must keep in mind social information processing [1] [8]. The human mind forms mental representations [5], which control every aspect of actions, from thoughts to motor movements. The information content of mental representations is called mental content. It is the basic theoretical concept in analyzing HTI issues from a human point of view [5]. Understanding how people mentally represent can be clarified by understanding the relevant mental content. Through teaching and training, people can find and use new mental content.

Ethical and moral information is a good example of mental content that is important in human HTI analysis. Ethical practices, rules, norms and tacit principles constitute mental content. If a person does not engage in truthful net communication practices, it is a value system, and it is mentally represented. In the opposite case, the information is also represented in the individual's mental content. Thus, explaining ethical information processing can be grounded in mental content. The prevailing mental content, with its emotional and cognitive aspects, explains how people act.

Consequently, investigations of ethical issues related to HTI can be approached from an information processing point of view. Thus researchers investigate how the human mind operates in different ways when people participate in different forms of life. From this point of view, it is also possible to reconsider classical problems of ethics such as Hume's guillotine [9], or the "is–ought to" problem.

Hume's problem is tricky in several senses when we think about how ethical AI life should be designed. First, it discusses the relationship between facts and values. This problem is important, as intelligent systems are fact processors. Second, his moral theories are firmly grounded in empiricist psychological thinking. Sentiments or emotions and reasons or cognitions are a central part of his thinking. As discussed in more detail below, resolving Hume's problem has important implications for the modern ethics of intelligent systems, their social use and technology design thinking.

## 2 Intelligent systems and Hume's guillotine

Technologies are developed and justified to help people achieve their goals and to improve the quality of human life or wellness [9]. The objective of a new technology is to make people's lives easier than before. Making something happen is a form of *emancipation*, which means expanding the possibilities of life. Historically, emancipation refers to breaking free from the social conditions that enslave people [1]. Life can be restricted by many kinds of necessities, difficulties, limitations and non-ideal living conditions that prevent people from increasing their happiness. But many of these problems can be solved or improved through technological advancements. Intelligent systems hold particular promise in their capacity to improve human life by emancipating people from routine tasks.

Intelligent technologies represent a new technological revolution, like stone-cutting tools, the printing press, steam technology, electricity, and nuclear energy generated new forms of work and social organization in the past [9]. Modern intelligent systems such as AI, autonomous systems and robots can carry out complex tasks that previously required information processing from a human mind. These emerging technological applications have revolutionary implications for the industrial processes, office automation, intelligent medicine, teaching, autonomous traffic systems and intelligent finance of the future [10].

In addition to fast routine processing logical inferences, intelligent systems can make decisions between alternative sense-making courses of action. They can even learn to make classifications of their own, so people cannot predict the information states that they can generate. Their capacity to engage in selective information processing makes it possible for modern AI-based systems to compare the values of different information states to select the one that is most fit for purpose.

Ethics introduces a very specific way to think about 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 based on the justification that they are more ethical than others. Thus, intelligent technologies can make operational decisions on ethical grounds. They can choose between different courses of action based on defined ethical principles. For example, intelligent systems can prioritize children over middle-aged people in making decisions about the order of medical operations.

Over 250 years ago David Hume identified an important problem in attempts to identify a relationship between facts and values: "It is impossible that the distinction between for all good and evil can be made by reason" [12]. This dilemma, known as Hume's guillotine or the "is-ought to problem", is central to the study of modern ethics. Hume's guillotine claims that one cannot determine how things *should be* based on how they *are*. When designing ethically intelligent machines, this is a relevant conceptual problem. One can justly ask: Can machines that process facts do so ethically, and if so, how is this possible?

The difference between how things are and how they should be plays a central role in ethical design thinking about how to improve things. Ethical design focuses on how to move from a given system of prevailing values to a system that improves the quality

of human life. Scientific knowledge about human information processing is vital to this work.

Real-life ethics involves studying how human information processes, emotions, cognitions and social information processing can be used to analyze the emergence of ethical norms, rules and actions. How it is possible for the human mind to create new ethical norms and practices? Of course, people have developed ethical norms such as laws and ethically regulated action patterns. One way to approach real-life ethics is to evaluate how different aspects of human information processes have been used to discuss the nature of ethics.

## 3 Emotions and cognitions

Emotive ethics or emotivism serve as a good starting point for the present analysis of the ethical relevance of information processes. In human information processing, emotions represent an evolutionarily more basic system of thinking than cognition. Emotional areas of the brain develop earlier than cognition, and especially higher-level cognition such as thinking.

Emotional ethics considers emotions to be fundamental components of ethical thinking. It was central to British empiricism. Smith [13] and Hume [12], for example, recognized the importance of emotional processes or passions and sentiments. In the last century many important researchers such as Moore [14] and Ayer [15] have also supported emotivism in different forms.

In information processing concepts, a key feature of emotions is valence, which refers to the positivity or negativity of emotions [7]. Pleasure and pain, good and bad, sorrow and joy, warmth and coldness are examples of opposite valences. In ethics and information processing, valence defines the goodness of actions and situations, and thus is essential to deciding how positive or negative actions and respective situations are from an individual's point of view.

The problem of Hume's guillotine arises from the difference between emotions and reason or cognition. Hume discussed the differences between reason and passion. He argued that the function of reason is to decide whether something is true or false. Emotions or passions with morals "produce or prevent actions." The two units are separate in the sense that truth and falsehood, i.e. reason, cannot dictate emotions. A typical consequence is the differentiation between the theoretical (i.e. reason based and practical) and the philosophical (i.e., passion dominated). Moreover, the distinction between an act that is morally good or bad cannot be made based solely on reason [12]. Hume views ethics as human mental activity that entails how people feel and reason – not simply as a system of rules.

Cognitive and human cognition refers to how people process knowledge [16]. Individuals take information from their environment, and store and manipulate it; in turn, it regulates their actions. Cognition registers actions and thoughts that have led people to a particular situation, and stores this information as memories. Thus, cognition provides mental representations of situations and the actions that have led to these situations.

Several ethics frameworks have grounded prior thinking about cognition. Typical examples are Kantian [17] and deontological ethics and Moses's ethics or Ten Commandments [18]. These directions present explicated norms of behavior that define how one should act to act correctly. All such norms and principles are expressed in the cognitive mind.

Emotional processes are closely linked to cognition. Emotional states are based on an individual's understanding of the current situation. If the situation is cognitively understood to be risky or threatening, the emotional states are constructed based on danger-related emotions, such as excitement, fear and courage. If positive cognitions dominate the situation, emotional states are characterized by relaxation, happiness, humor and benevolence. Before the situation-related emotional representation is constructed in the human mind, its cognitive content must be clear to the individual [7].

The psychological process of linking a situation's cognitive and emotional representations is called appraisal. Appraisal is a core process in the psychology of emotions [7], which is often defined as the representation of an individual's emotional significance, and the associated emotional value of cognitions and actions. Emotions associated with the use of technologies are relevant in the study of technology-related and AI ethics.

Cognitions provide cognitive aspects of ethical experiences in particular situations. Emotions provide evaluative information about these situations – for instance whether situations are pleasant or unpleasant, and good or bad for the person experiencing them. Thus, ethical experiences arise both from cognitions and emotions. The two systems encode different aspects of experiences and the respective mental representations.

People learn from experience to associate their actions with the situations the actions have led to. Based on these learned experiences, they encode rules of good conduct in interacting with technologies, including intelligent technologies. People learn to use them, which generates memory representations about the consequences of their actions and reasons why particular types of actions should be avoided or pursued, i.e., are the actions or duties allowed or forbidden. The representation of an action, its end situation and the emotional analysis of this situation can be called primary ethical representation.

## 4 Social discourse

Ethics is social because people are social. Aristotle's [19] concept of "Zoon Politicon" expresses this aspect of human nature effectively. In their social actions, people organize themselves into an infinite number of types of social circles including sports clubs, house societies, non-governmental organizations, states, schools, religious communities, campers, families, entrepreneurs and taxpayers. Any social group that organizes a participant's actions around some system of rule-following actions can be thought of as a *form of life* [1] [2].

Forms of life are organized systems of action in which individuals can participate, and ethics is essential to forming them. If one is Catholic, they typically participate in the ceremonies of the holy week. People follow the norms and traditions of the event.

In families, most people strive to take care of their children and speak with them about the way people should live. Such discourses belong to the family form of life.

Forms of life have rules, but these rules keep changing. A key mechanism of such changes is social discourse in its numerous forms [19]. Social discourse entails communicating individual ethical rules and norms. People feel that something causes pain and identify the mechanism of action that led to those unpleasant feelings. The discourses in different contexts give people the ability to create common norms within society, which in turn shape the forms of life.

Social discourse entails both free and normed forms. For example, discussions among friends are different from discourses in enterprise executive boards. Much social discourse now takes place via social media and other media. Even academic and political disputes on ethical issues can be seen as aspects of social discourse. Yet, these discourses create many types of actions that regulate ethical rules, such as company policies. How to express oneself in meetings and how to group in restaurants are typical examples of rules that regulate actions in organizational forms of life. Social discourse creates various social norms to guide how people act when participating in different forms of life.

Ethical discourse is in many ways beginning to define how people should act in different forms of life. It creates basic regulatory norms and values. However, societies are often regulated by laws. Of course, law making is an outcome of social discourse that is private as well as administrative or political. The forms of these discourses can vary from one society to another; democracies organize their discourses differently from oligarchies or dictatorships. Nevertheless, there are always groups of people who create new forms of life through thinking and discourse.

The social process of creating informal, tacit and formal regulatory rules and principles for different forms of life has been analyzed in detail in discourse ethics. A key issue is that the ideas are submitted for social discourse in different forms. Ideally, ideas are analyzed by assessing the argumentation. If arguments are valid, it is possible to continue norming. However, if they are no longer valid, e.g. historical changes have made them outdated, the rules should be replaced [19].

A very large “sea of social discourses” creates socially shared norms and renews them constantly. Social attitudes keep changing, social experiences are communicated to other people, and the discourse converges into systems of tacit and explicit norms and values. As a whole, the system of emotional valences, social analysis of related actions and action types, as well as social discourse formulate the ethical process. This process creates the values people follow in their everyday lives.

Individuals’ primary ethical schemas form the basis of social discourse. Through small and large, formal and informal discussions, people form their views about what are the most important and fundamental ethical experiences and respective rules. Discourse ethics has investigated this process [19].

In discourse ethics, representations are submitted to argumentative or foundational analysis. Each primary representation or ethical rule will be submitted to the foundational discourse. Any ethical rules that cannot be argumentatively supported will be rejected. The discourse itself has layers and sub-discourses. The main outcome is a

system of ethical concepts, rules and principles. The unification of emotional, cognitive and social analysis can be called an ethical process.

## 5 Ethical information process and process ethics

The ethical information process is the source of practical ethics in life: It creates values and norms. Research on ethical processes is valuable, as it creates a picture of a society's ethical thinking. Understanding the ethical process also makes it possible to solve the problem of Hume's guillotine. It appears to be a result of insufficient analysis of the relationship between people's minds and actions.

The analysis of ethical processes represents a specific approach to the study of ethics, which can be supported by its importance in designing an ethical world. For example, instead of representing external norms for the right kind of patient care, designers can work to understand how people are *really taken care of*, for example in units for senior citizens, and what norms they follow in their daily lives.

This type of empirical ethics is intimately connected to the analysis of ethical processes, but it has an important difference. The former moves the focus from academic discussions to life as people live it, which leads to the tacit and explicit development of a society's ethics, while the latter refers to the analysis of how norms are created. It is thus an empirical model of metaethical processes in real life. Westermarck [4] studied the norms and values of empirical ethics, while the analysis of ethical information processes presented here concentrates on the process of creating values. In ethical information processes, the creation of values and following them are both important. I refer to ethics based on the analysis of real-life value creation processes as "process ethics" to distinguish it from earlier approaches to ethics.

Value creation is important for design thinking. If researchers understand the value creation process, they can improve it by providing empirical information on different aspects of the process. This shift from reflective to active involvement and influence is vital in designing ethical AI processes.

Ethical information processes can help circumvent Hume's guillotine. Hume makes the fundamental (unsupported) assumption that emotions and cognitions are opposites in some sense, and that reason cannot affect sentiments or emotions. However, there is no support for such a conceptual differentiation in modern research on the mind. One cannot derive that the two concepts are opposites based on the fact they are different. In this case, they can complement each other.

Human actions are jointly regulated by both emotions and cognitions. They have different functions, and both are necessary. Emotions determine the goodness or badness of actions and attribute personal meaning to individuals [7], while cognitions analyze actions and consequential situations. Thus, the two faculties together can construct ethical experience and primary ethical schemas. Social discourse turns these primary ethical experiences and schemas into socially agreed rules and even laws. In this way they perform functions within relevant forms of life. Thus, Hume's guillotine is a pseudo problem that arises from a mistaken conceptualization of the mind and actions.

The next section considers what ethical information processes and their analysis can add to our understanding of ethical machines.

## 6 Weak and strong ethical intelligence

Improved computing speeds and the fast growth of data have made it possible to design technical artefacts with the ability to perform tasks that previously only people could carry out. Such machines are called intelligent systems as they execute tasks that demand intelligence from people. In addition to fast routine processing logical inferences, machines can decide 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 that intelligent systems can generate. Consequently, intelligent systems can select between different sense-making courses of action.

The capacity to engage in selective information processing makes it possible for modern AI and machine-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 machine actions intelligent. Similarly, ethical rules can be used as heuristics in selecting between different tasks.

For these reasons, one can speak of ethics typical to using intelligent technologies in two senses: (1) the ethical use of technical artefacts in society or (2) the development of systems with ethical capacities of some type. Here I investigate what it means to have ethical machines and technical artefacts.

Intelligent machines can be either weakly ethical (machines implement heuristics created by humans) or strongly ethical (machines can generate their own new ethical rules and principles). Hume's guillotine is easier to solve based on the former case. However, it is important to first ask how ethical information processing is possible, and then to evaluate how weak and strong ethical AI differ from each other.

The fall of Hume's guillotine paves the way for the development of strong ethical AI systems. It is possible to analyze data and see its connections to actions. Machines can also help determine if the resulting situations are emotionally pleasant or unpleasant. By combining the facts with emotional valence information concerning particular situations, machines can discover new primary values for social discourse. They can construct primary ethical schemas and thus develop stronger ethical AI. Human social discourse would be required to decide whether these new primary schemas are valid.

## 7 Ethics in designing intelligent systems

Intelligent technologies introduce a new element of human actions and forms of life. Intelligent systems will perform an increasingly larger share of actions. Such systems can process ethical information and carry out ethics-requiring tasks. For example, social services or migration offices must analyze masses of applications for services, but they have to decide which can be accepted at least partly on ethical grounds.

Ethical processes must thus be studied using human research methods and approaches, but they should also be designed and improved. Importantly, intelligent technologies can help us understand how to design ethical processes. Value-based design and ethically aligned design are examples of how design thinking can help create ethical social processes or forms of life.

A core problem associated with designing ethical processes is how to implement ethics in machines. Weak AI is not a difficult case. Ethical norms can be implemented in AI programs by defining ethics-requiring situations and their factual properties. Intelligent systems can extract key information from data, and associated ethical norms can be followed in actions. For example, an underage person applying for a driving license can be taken out of line to return a year later. Designers of ethical processes and forms of life can build recognition–association type action models with ethical content.

Strong AI in ethical processes is a more challenging case. In principle, such systems can suggest primary ethical schemas. Therefore, they have the capacity to develop strong AI. However, as primary ethical schemas are always accepted by social discourse, the primary schemas suggested by intelligent systems must also be subsumed under social discourse before their acceptance.

The border between weak and strong AI is not absolute, but systems can differ in their strength. The strength of an ethical AI system is based on its capacity to create new ethical norms without human process time involvement. First, it is possible by means of data analysis to study possible pain- or negative-valence-causing situations. For example, data mining can identify new factors that cause illnesses. Such research has existed for a long time. For example, Durkheim [10] found a link between religion, social discourse, and suicide, and a connection between smoking and lung cancer was found in the 1960s. There is no logical obstacle to using intelligent systems to find such associations. Thus, human-supported AI and data mining can be used to find novel factual grounds for new ways of behaving. This kind of ethical AI system is known as machine-supported AI.

Another possibility is to ask machines to recognize features that are known to cause emotionally negative experiences. Human responses to different types of situations would first be registered to classify them as emotionally negative. AI programmes could actively search for new combinations. The information found can be associated with actions that produce 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 ethical norms without human involvement to process based on their factual data. Machines can use different approaches to analyze emotional valences typical to some situations, and associate the results of this emotional analysis with the actions. They can even analyze general social attitudes in these situations to gradually increase the autonomy of ethical systems. But human involvement can be relatively direct in creating new ethical rules.

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 primarily 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 investigate how intelligent systems can follow ethical norms in their actions.

Thus ethical information processing can be conceived as a spectrum with weak and strong ethical AI at either end. Weak AI systems can recognize critical features in situations and apply given ethical rules in these situations. In such cases, ethics are just a human-implanted feature in a recognition action system.

Yet 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. Analyzing the ethical process also provides clues about how machines can be used to improve existing ethical processes and create new ones. Thus, strong ethical AI systems can collect data, associate it with situations, and link these situations to emotional valence and respective actions. Such systems could develop new ethical principles to follow.

Finally, the outcome of Hume's guillotine is the unnecessary and mistaken juxtaposition of emotions and cognitions. People associate emotions with cognitive plans and experience the outcome of a situation as either positive or negative. Based on their experience, they create primary ethical schemas and practices, which through social discourse become general ethical principles and even juridical laws.

Social norms can be implemented in machines, which can in turn be part of ethical information processes. Intelligent systems can recognize situations and be aware of ethical feelings associated with these situations – and thus make ethics-based decisions about their actions. Intelligent systems can also identify new types of properties in situations and determine whether they are emotionally positive or negative. Thus, designers can move from weak to strong AI. Yet, designers are still necessary, and ethical processes aided by intelligent systems are not created for systems but for people.

## References


1. Saariluoma, P., Cañas, J. Leikas, J.: Designing for life. Macmillan, London (2016).
2. Wittgenstein, L.: Philosophical investigations. Basil Blackwell, Oxford (1953).
3. Kant, I.: Kritik der reinen Vernunft. [The critique of pure reason]. Felix Meiner, Hamburg (1781/1976).
4. Westermarck, E.: The origins and development of moral ideas. MacMillan, London (1906).
5. Newell, A. & Simon, H. A.: Human Problem Solving. Prentice-Hall, Engelwood Cliffs, NJ: (1972).
6. Neisser, U.: Cognition and Reality. Freeman, San Francisco, CA, (1976).
7. Frijda, N. H.: *The Emotions.* Cambridge University Press, Cambridge (1986).
8. Moscovici, S.: Social representations. Polity Press, Cambridge (2000).
9. Bernal J. D. ; *Science in History*. Penguin books, Harmondsworth (1969).
10. Tegmark, M.: Life 3.0. Penguin Books, Harmondsworth (2017).
11. von Wright, G. H.: *Explanation and Understanding*. London: Routledge and Kegan Paul (1971).
12. Hume, D.; A treatise of human nature. Dent, London (1972).
13. Smith, A.: The wealth of nations. Dent, London (1975)
14. Moore, G. E.: Principia ethica. Cambridge University Press, Cambridge (1996).
15. Ayer, A.: Language, truth and logic. Victor Collanz, London (1936).
16. Anderson, J. R.: *Rules of the Mind*. Erlbaum, Hillsdale, NJ: (1993).
17. Malik, K.: A Quest for a Moral Compass. Atlantic Books, London (2014).
18. Habermas, J.: Diskursethik (Discourse ethics). Surkamp, Frankfurth am Main (2018).