6 Kant’s Argument from Moral Feelings: Why Practical Reason Cannot Be Artificial

Abstract: Can practical reason be artificial? The answer, from a Kantian point of view, is clearly negative: Practical reason cannot be artificial. After a preliminary remark on the possibility of Kantian moral machines (1.1) and some basics on the concept of practical reason (1.2) and Kant’s intuitionism (1.3), I will argue that in a Kantian model of moral obligation, the typical (human) moral subject has moral feelings and must have them in order to cognize the validity of the moral law as a categorical imperative (1.3). Using the knowledge argument against physicalism and functionalism, I shall argue that computers have no feelings and, a fortiori, no moral feelings; therefore, computers are not moral subjects (1.4). This conclusion is based on a Kantian I feel rather than I think (2.). I will then tackle two problems with this argument (3). I will conclude with an analogy (4): Just as planets do not fly, computers do not feel.

Artificial Intelligence (AI) has given rise to a variety of moral, juridical, economic and political, or for short, practical questions that need to be answered soon – from how to deal with self-driving cars to AI being the end of the human race due to some kind of technological singularity.¹ Since AI is, at least by philosophical standards, a fairly recent phenomenon, both these practical questions and problems as well as the possible answers and solutions to them are fairly new. Note, however, that these answers in turn will depend on foundations that are a far cry from unbiased or innocent; in applied ethics and political philosophy, one arrives very quickly at traditional questions and positions that one needs to discuss both on a metaethical and a normative level in order to provide sustainable answers. Hence, it is no surprise that in papers about

¹ This paper is a revised version of a paper published in Journal of AI Humanities, 2018, vol.2, 67–91 (‘Can practical reason be artificial?’). – I shall like to thank the organizers of The First International Conference on Artificial Intelligence Humanities held at Chung-Ang University, Seoul on August 16, 2018; my special thanks go to Prof. Chan Kyu Lee and Prof. Hyeongjoo Kim. I would also like to express my gratitude to Sorin Baiasu, Larissa Berger, Richard Evans, Markus Lohrey, Christian Prust, Elke Schmidt, and Thomas Sukopp for helpful discussions.
moral machines arguments are put forward on the grounds of good old fashioned utilitarianism, for instance.²

However, practical questions or the applied ethics of AI are not my concern here. Rather, the question that I shall address belongs essentially to the philosophy of mind: Can practical reason be artificial? Practical reason is best understood, I submit, as a genuine power to cognize and will the good. From a Kantian point of view, the answer to that question is clearly negative: *Practical reason cannot be artificial*. It is tempting to think that this answer has a foundation already in Kant’s epistemological, or — as Kant would put it — *theoretical* thought that reason is always someone’s reason, so that there is no thinking without someone who thinks or can always think *I think*. I will briefly look into this, but my focus will be on Kant’s practical philosophy. From this practical point of view, too, the conclusion that practical reason cannot be artificial is quick, solid and inevitable; for practical reason is free and computers are not. However, my approach is different; it is based on the idea that moral reason comes along with moral feelings that computers cannot have. After a preliminary remark on the possibility of Kantian moral machines (1.1) and some basics on the concept of practical reason (1.2) and Kant’s intuitionism (1.3), I will argue that in a Kantian model of moral obligation, the typical (human) moral subject has moral feelings and must have them in order to cognize the validity of the moral law as a categorical imperative (1.3). Using the so-called *knowledge argument* against physicalism and functionalism, I shall argue that computers have no feelings and, *a fortiori*, no moral feelings; therefore, computers are no moral subjects (1.4). This conclusion is based on a Kantian *I feel* rather than *I think* (1.5). I will then tackle two problems with this argument (2). I will conclude with an analogy (3): Just as planets do not fly, computers do not feel.

## 1 The Argument From Moral Feelings

It is easy to understand a Kantian argument that, if true, clearly rules out the possibility that computers have practical reason. It is the *argument from transcendental practical freedom*. Here is a very brief sketch: Moral obligation presupposes transcendental practical freedom of practical reason. Such freedom is, negatively speaking, independence from natural causality or physical determinism. Computers, however, are determined by the laws of physics; therefore, they cannot be free. But practical reason — and thus the human being — is free,
and it needs to be free for morality to make sense (Kant is an incompatibilist);\(^3\) therefore, no computer can have practical reason. Note that even on a non-deterministic understanding of physics, and even with regard to quantum computers, this argument from freedom will hold. For freedom is not only, negatively speaking, independence from natural causes; it is also, positively speaking, the faculty of determining oneself in an autonomous act of absolute spontaneity, and such spontaneity, unlike chance, is not lawless.

Thus, one can easily see that at least from a Kantian point of view it is quite obvious that a computer cannot have practical reason. Since this is so obvious, I would like to address or rather develop an argument from Kant’s practical philosophy that often goes unnoticed; I shall call it the argument from moral feelings.\(^4\)

1.1 A Preliminary Remark: Kantian Moral Machines

Alan Turing once listed a number of things that people think computers will never be able to do; these included the ability to “tell right from wrong”. Of course, it depends on what one means by ‘telling right from wrong’, but at least with regard to the output of such ‘telling’, that assumption might very well turn out not to be true. There has been a serious debate about “moral machines” for quite some time already,\(^5\) and the development of robots raises moral questions that are not only of theoretical (or philosophical) interest, so to speak, but are being addressed quite practically. Moral algorithms seem possible, and such algorithms may not only help, for instance, judges to make moral decisions, but will soon make, in some very limited sense, moral decisions on their own;\(^6\) just

---

\(^3\) Kant is an incompatibilist in the sense of denying the possibility of being a moral agent that is not strictly speaking free; morality presupposes absolute spontaneity (which is incompatible with determinism). Kant is a compatibilist only in that sense that freedom and determinism are compatible on the assumption – which is not the conceptual assumption in the debate about compatibilism and incompatibilism – of the difference between the noumenal and sensible world such that in some (transcendental or noumenal) respect the agent is free whereas in another (empirical or sensible) respect she is not; cf. Schönecker (2005).

\(^4\) In the English literature (with the famous exception of Antonio Damasio), it is common to distinguish between feelings and emotions; the former are understood as non-intentional (e.g. pain), the later as intentional (e.g. love). I will speak of feelings throughout without thereby referring to feelings in that strict sense. Moral feelings, in that sense, should be called moral emotions.


\(^6\) It is very difficult to put into language what computers or robots do – which is (almost) a case in point, because strictly speaking, computers and robots do not do anything, if by deeds we mean those acts that persons perform. As I see it, robots do not think or feel, they do not
think of so-called autonomous vehicles and the related trolley-problem. It is maybe tempting to assume that such moral machines must be based on some kind of utilitarian reasoning, given the mathematical character (and \textit{prima facie} easiness) of a utilitarian or hedonistic calculus. Given the \textit{formal} aspect of Kant’s famous categorical imperative and the idea of universalization, however, this could be a prejudice; a computer might be able to perform a moral algorithm on Kantian grounds as well. Recall the basic idea of the so-called natural law formula: Suppose someone has a maxim, for instance, that she will commit suicide when her life irreversibly brings about more suffering than agreeableness. The categorical imperative obligates her to ask herself whether such a maxim could be a universal (natural) law such that everyone who experiences more suffering than agreeableness will actually kill himself or herself; then she might realize that this leads to some kind of contradiction. There has been a long and ongoing debate on how to understand the contradiction Kant has in mind; but at least on a somewhat formal (logical) interpretation of the contradiction involved, a Kantian moral machine that runs a universalization test seems possible.

\section*{1.2 Kant’s Concept of Practical Reason}

‘Practical reason’ is (pure) good volition: “Every thing in nature works in accordance with laws. Only a rational being has the faculty to act \textit{in accordance with the representation} of laws, i.e. in accordance with principles, or a \textit{will}. Since for the derivation of actions from laws \textit{reason} is required, the will is nothing other than practical reason” (GMS: 412). It is important, however, to differentiate three aspects of Kant’s concept of practical reason or good will: The noumenally-good will, the practically-good will, and the holy will. The \textit{noumenally-good will} is the autonomous will that \textit{as such} wills the good. It is this noumenally good will Kant presupposes when he says that “a free will and a will under moral laws are the

---

\textit{play} a game nor do they \textit{make} moral decisions that would be sufficiently similar to what persons do, let alone truly perform these actions. So when I say that a robot ‘makes a moral decision on its own’, this is to be taken at best by way of analogy. I will get back to this later.


9 Cf. GMS: 427: “The will is thought as a faculty of determining itself to action \textit{in accord with the representation of certain laws}.”
same” (GMS: 447, emphasis mine). As a moral faculty, it gives the moral law (the categorical imperative) for imperfect beings and, by means of moral feelings, it is also a motivating force. Every human being has such a will, even if he or she acts immorally.

The noumenally-good will is the basis both for the practically-good will and for the holy will. The practically-good will is the will that finite beings have when their volition is indeed moral; it is the noumenally-good will considered as a will that manifests itself successfully in a finite being against the influence of inclinations and desires. For imperfect beings, to act morally (to act with a practically-good will) means to act from duty. The noumenally-good will that is manifest in a person without (active) sensual hindrances is what Kant calls the holy will; it only belongs to God and other holy beings. These beings have no inclinations and desires contrary to the good; the “will whose maxims necessarily harmonize with the laws of autonomy is a holy, absolutely good will” (GMS: 439). The noumenally-good will as such (regardless of its being incorporated in a finite being) cannot be differentiated from the holy will (regardless of its being incorporated in an infinite being). It is a noumenal causality: “The rational being counts himself as intelligence in the world of understanding, and merely as an efficient cause belonging to this world does it call its causality a will” (GMS: 453, first emphasis mine).

This will is then identified with the will that is autonomous, i.e., with autonomy itself: “if we think of ourselves as free, then we transport ourselves as members into the world of understanding and cognize the autonomy of the will, together with its consequence, morality” (GMS: 453, m. e.). Note how Kant continues: “... but if we think of ourselves as obligated by duty, then we consider ourselves as belonging to the world of sense and yet at the same time to the world of understanding” (GMS: 453, m. e.). Thus the free will is the noumenal will (pure practical reason), and autonomy is its property. In some contexts, this will is considered not as the will of a human being that is also part of the sensible world, but as a noumenal will only: as “a mere member of the world of understanding, all my actions would be perfectly in accord with the principle of the autonomy of the pure will” (GMS: 453, m. e.). It is important to keep in mind that the noumenally-good will as such is not only a mere capacity to act morally; for this will as such wills the good. Nonetheless, it is the noumenally-good will that enables the human being to act morally; thus, for the human being – who is a member both of the noumenal and of the sensible world – the

noumenally-good will is indeed a capacity. Also, unless autonomy and having a practically-good will are not the same, a scoundrel would not be autonomous – which he actually is insofar as even he, to some extent, wants to be morally good, i.e., insofar he has a noumenally-good will (we shall return to this later).\textsuperscript{12}

1.3 Kant’s Moral Intuitionism

It is often striking to see how defenders of strong AI\textsuperscript{13} find it obvious that computers “can do many things as well as or better than humans” (Russell & Norvig 2016, 1022, m. e.). But such an assumption, of course, is begging the question; for the point is rather whether they can do anything a human being can do when it comes to feeling, thinking, and acting. On Kant’s account, there is no ‘doing’ in any narrow sense here. Human actions, strictly speaking, are not only free actions; if they are guided by the moral law, they are embedded in moral feelings. Computers have no feelings; therefore, they do not act morally even if they make decisions in accordance with duty. Let us take a closer look at this argument from moral feelings.

To the present, Kant is believed to defend, as Edmund Husserl put it, “an extreme and almost absurd rationalism” (Husserl 1988: 412), an “extreme intellectualism” (ibid.) that leaves no room for feelings. Such a position betrays historical and textual ignorance. For even every beginner in an introductory class on Kant’s ethics will learn that Kant consistently argues that reason “obviously” (GMS: 460,9, m.H.) can only bring about actions by means of feelings; hence, feelings necessarily come into play here already as a determining (motivating) ground. However, it is important to realize that on Kant’s account, feelings serve a much more important function.

As we have already seen, Kant draws a very strict line between holy and non-holy beings. Whereas holy beings always will what a good will wants, this is not true for non-holy, sensuous-rational beings. For them, the moral law is always a categorical imperative that necessitates them. Let me quote Kant in more detail here:

\textbf{12} In the last section I have drawn from Schmidt & Schönecker (2018) and Schönecker & Wood (2015).

\textbf{13} By “strong AI” I mean for my purposes that a computer or robot could have consciousness, an inner life (qualia) and really think the way we do. Such a computer would not just imitate thinking, and would not just imitate moral thinking, but really think and therefore also think morally.
If reason determines the will without exception, then the actions of such a being, which are recognized as objectively necessary, are also subjectively necessary, i.e. the will is a faculty of choosing only that which reason, independently of inclination, recognizes as practically necessary, i.e. as good. But if reason for itself alone does not sufficiently determine the will, if the will is still subject to subjective conditions (to certain incentives) which do not always agree with the objective conditions, in a word, if the will is not in itself fully in accord with reason (as it actually is with human beings), then the actions which are objectively recognized as necessary are subjectively contingent, and the determination of such a will, in accord with objective laws, is necessitation, i.e. the relation of objective laws to a will which is not thoroughly good is represented as the determination of the will of a rational being through grounds of reason to which, however, this will in accordance with its nature is not necessarily obedient. The representation of an objective principle, insofar as it is necessitating for a will, is called a ‘command’ (of reason) and the formula of the command is called an imperative. All imperatives are expressed through an ought and thereby indicate the relation of an objective law of reason to a will which in its subjective constitution is not necessarily determined by that law (a necessitation).

The crucial step is to see that this necessitation is experienced by the feeling of respect (which in turn has a negative and a positive aspect that I cannot deal with here). But this feeling is not just a side effect, as it were. Since by ‘necessitation’ Kant means nothing but the fact that for non-holy, sensuous-rational beings like us the moral law is an imperative, that is, a duty, the obligation involved in this is experienced in the feeling of respect. As a matter of fact, it is not only somehow experienced, but cognized by this feeling: “What I immediately recognize as a law for me, I cognize with respect” (GMS: 402, fn.). And it is important to see that Kant’s famous theory of the ‘fact of reason’ is directly related to this thought. In § 7 of the Critique of Practical Reason, Kant formulates the categorical imperative; a bit later, he says that one could call the “consciousness of this fundamental law a fact of reason” (KpV: 31). The so-called factum-theory explains our insight into the binding character of the moral law; it is, among other things, a theory of justification. The basic idea is that there can be no deduction of the categorical imperative in any normal (deductive) sense, and yet the objective reality of the moral law is “nevertheless firmly established of itself” (KpV: 47, m. e.). In our consciousness of the categorical imperative, the moral law is immediately given in its unconditional and binding validity; in this sense (but only in this sense), the factum theory is a theory of moral self-evidence. This consciousness of the categorical imperative, however, is determined by the feeling of respect, that is, the unconditional validity of the categorical imperative is given in the feeling of respect.

---

15 “Hence the objective reality of the moral law cannot be proved by any deduction” (CPrR: 47,15).
Hence, it is through the feeling of respect that we cognize the validity or binding character of the moral law. Thus, Kant is by no means the pure rationalist that Husserl and others represent him as being. Rather, Kant is a moral intuitionist. A moral intuitionist is someone who holds the view that we cognize the validity of the moral law, the moral You ought, not by some kind of deductive (or inductive or abductive) reasoning, but by means of a certain kind of self-evidence, through a feeling. It is important not to misunderstand these claims: On Kant’s account, it is not the content of the categorical imperative that is understood through the feeling of respect; what we ought to do or omit we cognize by reason and some kind of universalization. Kant is not a moral sentimentalist. Also, the moral law itself does not depend on the moral feeling of respect for its validity; it is not that the moral law is valid because we have that feeling. Yet what we do cognize through the feeling of respect is that we ought to act morally, that the moral law is categorically binding.

1.4 A Kantian Knowledge Argument

From this, however, it follows that a computer can have no practical reason. To see this, we have to take a brief look at the so-called knowledge argument put forward in one version or another by Thomas Nagel and, historically more properly speaking, Frank Jackson. This is quite an intricate story, and here we can only sketch the main idea behind it.

For our purposes, recall Jackson’s thought-experiment about Mary: Think of her as a scientist who knows everything there is to know about colors and their perception from the point of view of the natural sciences; but Mary, being locked in a room with only black and white books, TVs, etc., has never seen

16 Colin & Varner & Zinser (2000, 260) see that emotions have more than just a motivational function; they do not, however, recognize their cognitive function as regards the validity of the moral law.

17 Cf. the pertinent texts by Nagel (1974) and Jackson (1982). Nagel’s argument is not that we can never understand or intuit or have access to someone else’s inner life and the phenomenal qualia involved. The point is that this understanding is subjective and not objective, i.e., it cannot be reached by a third person perspective and hence not by the natural sciences (cf. Nagel 1974, 441 f.). So in some sense, the bat-example is misleading. Even if there were only human beings and no other sentient organisms, the problem (or fact) Nagel describes would remain. Levine’s explanatory gap argument (1983) and Chalmer’s related ‘hard problem’ are somewhat similar, but still different, since here the focus is not so much on the physical inexplicability of the first-person-perspective as such but on the (presumably inexplicable) relation between consciousness (qualia, what-it-is-likeness) and their physical relatum as well as function.
any objects that are not black or white. One day, however, she leaves her room and actually sees something that is, say, red. Now according to physicalism (materialism, naturalism), only natural (physical) objects exist, and natural objects are those described and explained by physics (and possibly by chemistry, biology, or neuroscience). If this were true, then Mary would not gain any new knowledge of a quality she had not already known because everything there is to know about colors from an scientific, objective, third-person perspective, she already knows. But there is something she did not know before she left the room, to wit, how it feels or what it is like (the later is Nagel’s famous formula, of course) to see something red, to experience a certain quale; therefore, there is something in the world that is not physical, i.e., not fully describable by physics. This something is consciousness making phenomenal experiences. Thus we could know everything there is to know about the physical or functional facts concerning a mental state (such as having a perception of something being red) and still we would not know everything about that mental state; therefore, this mental state cannot be identical with or be reduced to those physical or functional facts. From this it follows, some at least have argued, that physicalism is false.

Although there is some dispute as to which mental states are qualia or are accompanied by qualia, it is obvious that feelings are indeed qualia. But then the argument is apparent: We can know everything a computer is made of and how it works. There is no what it is like to be a computer, and therefore, unlike beings for whom there is a certain phenomenal inner life, to be in a computational state is not to be in a mental state (and vice versa). And so unless computers experience qualia, they cannot have practical reason. For practical reason comes along with practical necessitation through the feeling of respect; the categorical imperative cannot be understood without this feeling; since computers have no feelings, and a fortiori no feeling of respect, they cannot understand the categorical imperative.

That is the basic argument. From a Kantian point of view, there are three more important points: First and only in passing, I should note that in his later work (The Metaphysics of Morals) Kant further developed his theory of moral feelings by distinguishing four kinds of moral predispositions and, consequently, four moral feelings: the moral feeling proper, conscience, love of human beings as amor complacentiae, and self-respect.18 With regard to each of

18 Cf. Schönecker (2010). – Kriegel & Timmons (2021) claim to work on a ‘phenomenology of Kantian respect for persons’. For some reason or other, however, they pay no attention to ‘respect’ (as self-respect) as one of the four moral predispositions. They also seem to be unaware
these feelings Kant stresses that there is no obligation to have them; for to have these feelings is already a necessary presupposition to make sense of the very concept of duty in the first place. Second, to Kant, practical reason is the noumenal will that both cognizes and wills the good; it is autonomous and therefore a noumenal causality. As I have indicated already, this is a complicated story, but moral feelings cannot be naturalized as they are brought about by reason which cannot be naturalized; so even if computers did have feelings, they could not have the feeling of respect, because this feeling has its source in reason which is not a natural (physical) entity. Third, Kant, too, understands feelings as qualia. Of course, Kant did not use this term. Nevertheless, he had a clear understanding of the fact that feelings have a phenomenal side that cannot be grasped by physical knowledge but must be experienced. The phenomenal side of feelings is particularly emphasized by Kant in his theory of beauty.19

Feelings as such, says Kant in the so-called First Introduction to the Critique of the Power of Judgment, “cannot be explained at all”; rather, they “must be felt, not understood [eingesehen]” (EEKU: 232).20 In a similar vein, Kant writes in the Metaphysics of Morals that “pleasure and displeasure cannot be explained for themselves” (MS: 212).

In any case, the purity in pure practical reason by no means suggests that there are no feelings involved in the process of moral self-determination. The purity of practical reason consists in its being free from considerations of happiness and self-love; in human beings, it is pure reason that becomes practical on the strength of moral feelings. So even if a computer ‘makes a decision’ (as it were) on the basis of a moral algorithm, it has no idea of what it is ‘doing’ (as it were): it has no understanding whatsoever of what the moral law as a categorical imperative really is. In Kant’s terminology, a computer can perform actions (as it were) according to duty. But it certainly cannot perform actions from duty. And it certainly has no conscience or self-respect; as I see it, such a claim is not even apprehensible.

Before I move on to two problems for this Kantian position, let us have a quick look at a possible further argument, the argument from the power of judgment; given the latitude of many ethical duties, this is obviously an important

---

20 In German: “Man sieht hier leicht, daß Lust oder Unlust, weil sie keine Erkenntnisarten sind, für sich selbst gar nicht können erklärt werden, und *gefühl*, nicht eingesehen werden wollen”.

aspect of practical reason. The argument could run like this: Following Kant, the power of judgment is the “faculty of thinking the particular as contained under the universal“ (KU: 179). If there is a rule, then the faculty of the power of judgment is the “faculty of subsuming” something particular under this rule; Kant calls this the determining power of judgment (*bestimmende Urteilskraft*). If the rule is yet to be found for something particular that cannot be subsumed under an already existing rule, then Kant calls it reflective power of judgment (*reflektierende Urteilskraft*). At least for the determining power of judgment, Kant argues, there can be no further rule. For if one “wanted to show generally how one ought to subsume under these rules, i. e., distinguish whether something stands under them or not, this could not happen except once again through a rule. But just because this is a rule, it would demand another instruction for the power of judgment, and so it becomes clear that although the understanding is certainly capable of being instructed and equipped through rules, the power of judgment is a special talent that cannot be taught but only practiced” (KrV: A133/B172). Put another way: There can be meta-rules on how and when to apply rules; but on pain of a vicious circle or an infinite number of rules, there must be a point at which the power of judgment takes action without applying a rule. Computers, however, have nothing but rules to work with, i. e., nothing but algorithms (and data, of course, in regard to which they are applied). If the power of judgment is a faculty that does not follow rules, then this faculty cannot be something a computer could have. The “lack of the power of judgement,” says Kant, “is that which is properly called stupidity” (ibid.); in this sense, computers are stupid.

Yet it is dubitable whether this argument actually goes through. The necessity of something like a power of judgment is due to the fact that there is no complete ascertainment or definition of all possible concepts and cases *a priori* or in advance. But if a decision based on the power of judgment is not based on a rule, on what is it based? It had better not be based on chance; for that is something a computer could do (following the rule to choose randomly). One might think that the power of judgment has to do with something like intuitions; but intuitions (the way G. E. Moore, for instance, understands them) are different and have nothing to do with the power of judgment. However, intuitions broadly

---

21 Kant actually thinks that all ethical duties (unlike juridical duties) are wide duties. Note, however, that there are negative ethical duties that despite being duties of omission have latitude; cf. Schmidt & Schönecker (2018).

22 Cf. KrV: B 756: “One makes use of certain marks only as long as they are sufficient for making distinctions; new observations, however, take some away and add some, and therefore the concept never remains within secure boundaries”.

understood as somewhat unconscious, strong, quick seeing(judging)-as-true states of mind (as understood in moral psychology), might very well not be random, but based on some (unconscious, strong, quick) weighing of goods, and such a weighing could follow rules. In any case, I would not be ready to defend the argument from the power of judgment.

2 I Think vs. I Feel

Obviously, moral subjects will something, and they act on the basis of their volitions. But they also think, and even if it is disputed how and how much thinking is actually involved in moral decisions as such, it cannot be disputed that moral subjects must also think, at least with regard to the cognition of the surrounding world in which they will and act; moral knowledge involves non-moral knowledge about the internal and external world. Now Kant famously argues (in the second edition of the Critique of Pure Reason) – or appears to argue, anyway – that there is no thinking and hence no knowledge of the internal and external world without self-consciousness; the I think, he says, “must be able to accompany all my representations” (KrV: B 131). What exactly this means has been the object of painful and long discussions among Kant-scholars.23 One problem is the nature and relation of what Kant calls the analytic unity of apperception to what he calls the synthetic unity of apperception (cf. KrV B133 ff.); it is neither clear what exactly each of these unities really is and in what sense the synthetic unity of apperception is a necessary condition of the analytic unity of apperception. Very generally speaking, I take the basic idea to be that all thinking as the “capacity to judge” (A69/B94) involves the synthesis24 of representations as someone’s representations in a judgment, such that these representations – as well as the act of synthesizing them – belong, and must belong,25 to a self-conscious I that can always say I think (these representations). In any case, if Kant is right, and if it is true that a computer has no I, then a computer does not and cannot think, and it is not intelligent the way human beings are.26 At best, then, a computer (AI) can only imitate intelligence and moral thinking.

24 Cf. KrV: A69/B94: “All judgments are therefore functions of unity among our representations”; categories are “concepts of synthesis” (A80/B106). Evans (2022, 50) speaks of containment, comparison and inherence as the three basic operations.
25 Friebe (2005, 53) seems to think that such a reading is too strong.
26 For further discussion of this argument, see Baiasu (2022) and Evans (2022) in this volume.
I’m not quite convinced of this argument. It seems true to me that (developed) human beings must always be able to think I think when they think (not that they always do think I think, of course). Now, if this is true, and if it is true that computers have no I (which I think is true indeed), then computers do not think the way we do. Still it might be justified to say that they think: If theoretical (not practical) thinking is essentially the act of synthesizing content that requires a center or unit by which this act is performed, then one could possibly understand this center or unit as the control unit of a computer. My point is that the act of combining content (synthesis) might require neither Meinigkeit nor self-consciousness (apperception, as Kant calls it) as a higher (egological) form of Meinigkeit, but just a control unit. This unit must not necessarily be aware of itself because the content of (something like) perceptions (intuitions) and (something like) thoughts or propositions does not require an I that thinks it and not even a form of Meinigkeit; all that is needed for thinking as synthesizing is synthesis itself, not synthesis that is aware of itself.27

It seems to me that Kant himself might have been aware of this and that all he claims is that the way a human being thinks is such that an underlying I is involved that thinks. Now in thinking I think a human being is not only aware of itself as the I that thinks I think and only this (transcendental apperception, mere spontaneity), but also of itself as determinable by representations in the order of time because “without any empirical representation, which provides the material for thinking, the act I think would not take place” (KrV: B 423, fn.).28 So one’s own existence (Dasein) is given in two ways: as the existence of the mere and “wholly empty representation: I” (KrV: A345/B404), and as this very (numerically identical) I that thinks something (has representations in time) and therefore thinks itself as determinable in time and therefore also exists as an empirical I. This is why Kant calls the I think an “empirical proposition” (KrV: B422 fn, cf. B 404, 420, 428) and why he also says, in direct connection with this, that the proposition I think

27 There remains, of course, the question of how computers have representations. Friebe (2005, 61) is quite right in claiming that representations (Vorstellungen) as such are always someone’s representations; and it might also be true that the property of being my representations (Meinigkeit) does imply an I whose representations they are. However, the question remains whether thinking must be understood as an operation that involves representations as something mental.

28 Cf. KrV: B 420: “But because my existence in the first proposition [I think, D. S.] is considered as given, since it does not say that every thinking being exists (which would at the same time predicate absolute necessity of them, and hence say too much), but only ‘I exist thinking,’ that proposition is empirical, and contains the determinability of my existence merely in regard to my representations in time.”
“contains within itself the proposition ‘I exist’” (KrV: B 420, fn.): As an empirical proposition, the I think contains the potentially (determinable) empirical existence of the I.29

In this context, Kant repeatedly criticizes Descartes for (allegedly) saying “Everything that thinks, exists” (KrV: B422, fn.; cf. B 404 and B420). If this were true, he argues, “the property of thinking would make all beings possessing it into necessary beings” (KrV: B422, fn; cf. B420). Thus, Kant criticizes the following argument (which he claims is Descartes‘): Everything that thinks, exists; I think; therefore, I exist. He does so by criticizing the first premise. Taken literally, it seems almost impossible to understand what his critique of the first premise (‘necessary beings’) really consists in. However, Kant’s argument does make sense if we assume that not every act of thinking is accompanied by the I think. ‘Everything’ that thinks and does so by also thinking I think is “given” (KrV: B 420) its existence; and if it were true that every act of thinking involved an I that thinks, then the major premise of Descartes’ inference would be true as well. If, on the other hand, it is possible to think of thinking as not involving an I that thinks, then obviously I exist is not implied by that thinking (it is only implied, as Kant says, by the I think).30

In the introductory passage of the chapter on the Paralogism of Pure Reason, Kant makes the following remark:31

But right at the start it must seem strange that the condition under which I think in general [daß die Bedingung, unter der ich überhaupt denke], and which is therefore merely a property of my subject, is at the same time to be valid for everything that thinks, and that on an empirical-seeming proposition we can presume to ground an apodictic and universal judgment, namely, that everything that thinks is constituted as the claim of self-consciousness asserts of me. But the cause of this lies in the fact that we must necessarily ascribe to things a priori all the properties that constitute the conditions under which alone we think them. Now I cannot have the least representation of a thinking being through an external experience, but only through self-consciousness. Thus such objects are nothing further than the transference of this consciousness of mine to other things, which can be represented as thinking beings only in this way. (KrV: A 346 f./B 404 f.)

29 Though this is an intricate story; cf. Kim (2017).
30 I do not mean to say that the alternative to an egological account of the self in self-consciousness is a non-egological account, because in the later account there is consciousness, too (though without an I), that is, conscious of itself (for this difference cf. Frank, 2015, 14 ff.). I mean thinking without consciousness; it allows for self-knowledge, though.
31 In the earlier version of this paper (s. fn. 1) I misinterpreted this passage.
So two things ‘seem strange’: First, that the *I think*\(^{32}\) as ‘the condition under which *I* think in general’ and which is ‘merely a property of my subject’, shall be a condition ‘valid for everything that thinks’; and, as a consequence, that that proposition (*I think*), which (as we have seen above) also has an empirical aspect, could be the ground for the proposition ‘that everything that thinks exists’ – which is exactly the Cartesian proposition Kant later vigorously opposes. And since we know for sure from the other passages that Kant indeed finds grave fault with Descartes’ argument (or rather its first premise), what ‘seems strange’ here must not be read as being only ‘seemingly’ strange, but as *really* being strange.\(^{33}\)

So maybe there is thinking without a self-conscious *I* that thinks and even without any consciousness. When it comes to feelings, however, we *necessarily* enter into a different world – the inner world.\(^{34}\) One can identify the activity of thinking\(^{35}\) and still leave it an open question of whether thoughts are thought by an *I* that thinks. This is impossible as regards feelings. There can be no feeling without *someone* who feels and for whom *it is like*.\(^{36}\) To say that there is a state which we could preliminary describe as “there is a feeling”, requires that there is an instance for whom it is to be in that state (of feeling something). Whatever

\(^{32}\) Throughout those introductory passages, the *I think* is discussed not only as a transcendental concept (cf. KrV: A341/B399) but also as the “sole text of rational psychology” (KrV: A343/B401).

\(^{33}\) Note that Kant says that “right at the start it must seem strange . . .” (the German original is even stronger, it seems to me: “Es muß aber gleich anfangs befremdlich scheinen . . .”). If it only said “Es muß aber anfangs befremdlich scheinen” (i.e. without the “gleich”) one could maybe get the impression as if Kant were to say that ‘at the start it must seem strange’, but later it doesn’t; but really it is later that Kant’s critique of Descartes becomes evident.

\(^{34}\) If cognitive acts have a phenomenality of their own (i. e. if there is a what-it-is-like to think), then these acts, too, require someone for whom it is-like to think (cf. Bayne/Montague, 2011).

\(^{35}\) Kant often speaks of the “actions of the understanding” (*Handlungen des Verstandes*, e. g. KrV: A69/B94, m. e.)

\(^{36}\) As Chalmers (1996, 147) puts it: “all it is for something to be in pain is for it to feel like pain. There is no distinction between pain and painful stuff, in the way there is a distinction between water and watery stuff. One could have something that felt like water without it being water, but one could not have something that felt like pain without it being pain. Pain’s feel is essential to it”. In her paper on computational models of emotion, Gu discusses so-called affective computing, but points out that “both emotion recognition and expression belong to the outer layer” of emotion – i.e. to “cognitive, behavioral, and physiological processes” (Gu 2010, p. 436) – and that the ‘core affect’ part has not yet been included in the discussion” (ibid. p. 440) on computational models of emotion. This “‘core’ emotion, often equated with the term *affect*, is the subjective *feeling* aspect of emotion” (ibid., p. 436).
feelings are, too, (bodily sensations, judgments, perceptions, evaluations), it is essential that one feels them. A feeling that is not felt is not a feeling; but for it to be felt there must be someone who feels it.37

3 Two Problems for the Argument from Moral Feelings

The argument from moral feelings is strong, I submit; still, it has two problems. First, what about holy wills? As we have already seen, it is a very important element of Kant’s ethics to distinguish between holy and non-holy beings. For the latter, the moral law is a categorical imperative and thus duty and obligation; for the former, there are no hindrances for morality to overcome, they have a perfectly good will. So must holy beings then be understood as moral machines? And would it not be true for a computer that always follows moral algorithms that it has such a perfectly good will? Well, no. It is certainly correct that a holy being cannot act from duty.38 But unlike computers, holy beings have a will; computers have no will; a fortiori, they do not will anything for the sake of the moral law. The claim that computers have no will is underpinned by the claim that volitions are intentional; computational states, however, are not intentional states;39 therefore, computers have no will. Still the question of what it means for a holy will to will and to act for the sake of the moral law without intermediary moral feelings is not easy to answer.

The second problem with the argument from moral feelings simply is this: Can computers really have no feelings? This, too, is a long and complicated story, and I can only sketch the problem and the possible solution. So here is the argument that maybe computers could or even do, after all, have feelings: We know that we have consciousness and feelings. At the end of the day, we have no clear, let alone comprehensive and convincing story to tell how this happens, how it really can be that we have such an inner life. However, provided that some kind of naturalistic evolutionary theory is correct, we do know that our ability to have mental events has developed out of unconscious matter. But if it is possible that the mind and its mental states somehow evolved out of matter, i. e., out of the brain and its embodiment – and possible it is on the

37 This does not rule out that feelings are also dispositional and as mere dispositions are not felt.
38 It also cannot have the four moral predispositions.
39 All of this is highly disputed, of course; recall the long and ongoing debate about Searle’s Chinese room argument. For a brief summary and critique cf. Gabriel (2018, 95 ff.).
assumption that mental states are real, regardless of whether they can be reduced to brain states or not –, then it could very well be possible that the mind and its mental states somehow could evolve out of a computer as yet another complicated assembly of matter. – Fair enough, I would say. So yes, that is possible if it is possible that the mind evolved (and evolves) out of matter. But from all that we know it is also very unlikely: A single biological cell is already extremely complex, let alone the brain, and complexity, as far as we know, is a precondition for, as well as a sign of, consciousness and feelings. By comparison, a computer is a very primitive object; there is no more reason to think that it has a mind than to think a sewing machine has one.  

4 Conclusion: Swimming Submarines, Flying Planets

To conclude, let me come back to Kant’s I think. The computer scientist Edsger Dijkstra famously argued that “the question of whether Machines Can Think . . . is about as relevant as the question of whether Submarines Can Swim,” or, one might add, whether airplanes can fly. His point, I take it, is that submarines of course can swim, i.e., move through water, despite the fact that they do not swim like fish, and that airplanes can fly despite the fact that they move through air without moving wings up and down or whatever. Following Dijkstra’s analogy, it seems sensible to hold that computers think despite the fact that they do not think the way we think; there is no I, and still they think. But this analogy between computers thinking and submarines swimming or airplanes flying is misleading. As always, it all depends on how one defines such words as “swimming”, “flying”, or “thinking”. If one defines “swimming” as “moving through water using limbs, fins, or tails”, then submarines do not swim; but why define “swimming” like this? To define a word – or to explicate what it is for a thing to be what it is – somehow presupposes a recognition of what is essential to that thing; for this, however, one needs paradigmatic cases. 

40 One could also, by the way, turn the tables and argue that the existence of mental states proves that something is wrong with Darwinist evolutionary theories. Thomas Nagel, among others, has done so recently; cf. Nagel (2012). For a critical view of Nagel’s moral realism, cf. Schmidt (2018).


computers on the basis of certain inputs perform certain operations that bring about certain outputs; and taking into account just the output – calculating, playing chess, driving a car, writing music – one is tempted to believe that computers think (and also that human beings think the way computers think). The question of how and by what means swimming is performed is not crucial to the concept of swimming. But the difference between a being that thinks I think, or at least between one that feels I feel, or that experiences I want, and a machine that has no such self-consciousness, is so enormous that those terms (to think, to feel, to want) should not be used for beings that have no I. To say that a computer feels is like saying that a planet flies just because it moves through space.

References

Kant’s moral writings (except for the Groundwork; here I use Allen Wood’s translation) are cited from the translation by Mary J. Gregor in the volume Practical Philosophy (The Cambridge Edition of the Works of Immanuel Kant), Cambridge 1996. The KrV is quoted according to the translation provided by Paul Guyer and Allen W. Wood (Cambridge Edition). All page and line numbers refer to the pagination of the Academy Edition.

EEKU Erste Einleitung in die Kritik der Urteilskraft/First Introduction to the Critique of Judgment
GMS Grundlegung zur Metaphysik der Sitten/Groundwork of the Metaphysics of Morals
KpV Kritik der praktischen Vernunft/Critique of Practical Reason
KrV Kritik der reinen Vernunft/Critique of Pure Reason
KU Kritik der Urteilskraft/Critique of Judgment
MS Die Metaphysik der Sitten/Metaphysics of Morals


