Journal of Philosophical Investigations
Print ISSN: 2251-7960 Online ISSN: 2423-4419
Homepage: https://philosophy.tabrizu.ac.ir

"It's a Human Thing. You Wouldn't Understand": Computing Machinery and Affective Intelligence

Robert Hanna

Professor, Department of Philosophy, University of Colorado, Colorado, USA. E-mail: contemporarykantianphilosophy@gmail.com

Article Info

Article type:
Research Article

Article history:

Received 17 September 2026
Received in revised form 29 September 2026
Accepted 6 October 2026
Published online 25 October 2026

Keywords:

Artificial intelligence; AI; digital technology; Turing Test; Affective intelligence; Immanuel Kant; Authenticity.

ABSTRACT

The strong thesis of artificial intelligence, aka strong AI, is that the mechanical powers of digital computing systems or digital technology more generally, can equal or exceed the achievements of rational human mindedness or human intelligence. In a recent book, Digital Technology for Humans, I've argued that not only is it not the case that The Turing Test is a sufficient condition for having rational human mindedness or human intelligence, but also that the very idea of artificial intelligence itself is nothing but a myth, since all digital computing systems, and all digital technology more generally, cannot, not even in principle, be intelligent in the sense in which rational human minded animals are intelligent, even when we also fully allow for our finitude, fallibility, and thoroughgoing normative imperfection. In this essay, however, I want to argue more specifically that our essentially embodied innate mental capacity for affective intelligence necessarily transcends the mechanical powers of any and all actual or really possible digital computing systems or digital technology more generally, even when we also fully allow for our finitude, fallibility, and thoroughgoing normative imperfection. By "affective intelligence," I mean intelligence that flows from our essentially embodied innate mental capacity for caring (i.e., desiring, emoting, and feeling).

Cite this article: Hanna, R. (2026). "It's a Human Thing. You Wouldn't Understand": Computing Machinery and Affective Intelligence. *Journal of Philosophical Investigations*, 20(56), 29-56. https://doi.org/10.22034/jpiut.2026.74804.4765

© The Author(s). Publisher: University of Tabriz.

"It's a human thing. You wouldn't understand."
(Dialogue and still frame from "I, Robot" (2004, dir. Alex Proyas)

Dasein's Being reveals itself as care. (Heidegger, 1962, 227)

Caring, insofar as it consists in guiding oneself along a distinctive course or in a particular manner, presupposes both agency and selfconsciousness. It is a matter of being active in a certain way, and the activity is essentially a reflexive one. This is not exactly because the agent, in guiding his own behavior, necessarily does something to himself. Rather, it is more nearly because he does something with himself. (Frankfurt, 1988c, 83)

Intruduction

Alan Turing's famous essay, "Computing Machinery and Intelligence," begins as follows:

> I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the term's "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. The new form of the

> problem can be described in terms of a game which we call the "imitation game" (Turing, 1950, 433).

In other words, Turing is using "intelligence" as a synonym for "the capacity to think"; and what he calls "the imitation game" is the same as The Turing Test. Correspondingly, the Stanford Encyclopedia article on "The Turing Test" glosses Turing's argument as follows:

> The phrase "The Turing Test" is most properly used to refer to a proposal made by Turing (1950) as a way of dealing with the question whether machines can think. According to Turing, the question whether machines can think is itself "too meaningless" to deserve discussion…. However, if we consider the more precise—and somehow related—question whether a digital computer can do well in a certain kind of game that Turing describes ("The Imitation Game"), then—at least in Turing's eyes—we do have a question that admits of precise discussion. Moreover, as we shall see, Turing himself thought that it would not be too long before we did have digital computers that could "do well" in the Imitation Game. The phrase "The Turing Test" is sometimes used more generally to refer to some kinds of behavioural tests for the presence of mind, or thought, or intelligence in putatively minded entities (Oppy and Dowe, 2021).

Reduced to its essentials, The Turing Test deploys written natural language, in the form of questions and answers exhanged between two writers, one of whom is a rational human experimenter and other a test subject—whether a machine or another rational human animal—as a criterion for the presence or absence of what I'll call "rational animal mindedness" or more simply, "intelligence," in the test subject. By "rational animal mindedness" or "intelligence," I mean the possession of a unified set of essentially embodied innate mental capacities, faculties, or powers for (i) *consciousness*, i.e., subjective experience, (ii) *self-consciousness*, i.e., consciousness of one's own consciousness, second-order consciousness, (iii) *caring*, i.e., desiring, emoting, or feeling, (iv) *sensible cognition*, i.e., sense-perception, memory, or imagination, (v) *intellectual cognition*, i.e., conceptualizing, believing, judging, or inferring, (vi) *volition*, i.e., deciding, choosing, or willing, and (vii) *free agency*, i.e., free will and practical agency. In specifically human animals, the unified set of these capacities constitutes our *rational human mindedness* or *human intelligence*, which is the same as our *human real personhood* (Hanna, 2018a: chs. 6-7). By "human" in this context, I mean not merely being a member of the biological species *Homo sapiens*, but also what Nietzsche meant by "human, all-too-human," namely, "finite, fallible, and thoroughly normatively imperfect." "Rational animal mindedness," as I am understanding it, applies to beings that aren't necessarily *biologically* human, since non-human rational minded animals in alien species, or other earth-

bound species, are wholly meaningfully conceivable, as fantasy writing and science fiction clearly show. So, what's being tested-for in The Turing Test, is an ability to use written natural language in such a way as to convince an interlocutor who already possesses rational human mindedness or human intelligence, that the other interlocutor also possesses all of the essentially embodied innate mental capacities that are jointly constitutive of rational animal mindedness or intelligence in human animals, namely (as per the above), consciousness, self-consciousness, caring, sensible cognition, intellectual cognition, volition, object-directed and act-directed intentionality more generally, and free agency.

Now, Turing himself apparently holds that something's passing The Turing Test in any given case is a sufficient condition for that thing's having rational human mindedness or human intelligence in that case. But that's obviously a mistake. Just because a machine can be created that emits meaningful written text in such a way that it convinces someone that it has rational human mindedness or human intelligence, it doesn't follow that the machine *actually has* rational human mindedness or human intelligence. Similarly, just because a machine can be created that convinces someone that it's God, or Satan, or The Wizard of Oz, it doesn't follow that the machine *actually is* God, or Satan, or The Wizard of Oz. Most people are naturally fairly gullible and trusting (I'm *not* saying that this is necessarily a *bad* thing, in fact I think it's often and even on the whole a very *good* thing—just that it's a fact), especially romantically-minded people, and therefore it's fairly easy to deceive them, and convince them to believe things that are actually false or bad, as all bullshitters, dictators, would-be dictators, other demagogues, con men, other grifters, advertising-copy writers, influencers, and other sophists know full well; and this of course also applies directly to The Turing Test. Consider, for example, the June 2022 *Google LaMDA* controversy:

> On June 11, 2022, The Washington Post reported that Google engineer Blake Lemoine had been placed on paid administrative leave after Lemoine told company executives Blaise Agüera y Arcas and Jen Gennai that LaMDA [Language Model for Dialogue Applications] had become sentient. Lemoine came to this conclusion after the chatbot made questionable responses to questions regarding self-identity, moral values, religion, and Isaac Asimov's Three Laws of Robotics. Google refuted these claims, insisting that there was substantial evidence to indicate that LaMDA was not sentient. In an interview with *Wired*, Lemoine reiterated his claims that LaMDA was "a person" as dictated by the Thirteenth Amendment to the U.S. Constitution, comparing it to an "alien intelligence of terrestrial origin." He further revealed that he had been dismissed by Google after he hired an attorney on LaMDA's behalf after the chatbot requested that Lemoine do so. On July 22, Google fired Lemoine, asserting that Blake had violated their policies "to safeguard product

> information" and rejected his claims as "wholly unfounded." Internal controversy instigated by the incident prompted Google executives to decide against releasing LaMDA to the public, which it had previously been considering. Lemoine's claims were widely pushed back by the scientific community (Wikipedia, 2026).

Clearly, according to Blake Lemoine, LaMDA passed The Turing Test, yet Lemoine was simply deceived by LaMDA. Therefore, passing The Turing Test in any given case isn't a sufficient condition for having rational human mindedness or human intelligence.

The strong thesis of artificial intelligence, aka *strong AI*, is that the mechanical powers of digital computing systems or digital technology more generally, can equal or exceed the achievements of rational human mindedness or human intelligence. In a recent book, *Digital Technology for Humans*, I've argued that not only is it *not* the case that The Turing Test is a sufficient condition for having rational human mindedness or human intelligence, but also *that the very idea of artificial intelligence itself is nothing but a myth*, since all digital computing systems, and all digital technology more generally, cannot, not even in principle, be intelligent in the sense in which rational human minded animals are intelligent, *even when* we also fully allow for our finitude, fallibility, and thoroughgoing normative imperfection (Hanna, 2025: ch. 1). In this essay, however, I want to argue more specifically that our essentially embodied innate mental capacity for *affective intelligence* necessarily transcends the mechanical powers of any and all actual or really possible digital computing systems or digital technology more generally, even when we also fully allow for our finitude, fallibility, and thoroughgoing normative imperfection

By *affective intelligence*, I mean intelligence that flows from our essentially embodied innate mental capacity for *caring*. As per the second and third epigraphs for this essay, I've appropriated the term "caring" from Martin Heidegger and Harry Frankfurt (Heidegger, 1962; Frankfurt, 1988c; see also Hanna and Maiese, 2009: ch. 5). And more specifically, as I noted above, by the term "caring" I mean our essentially embodied innate mental capacity for *desiring*, *emoting*, and *feeling*. Obviously, there are in fact many different modes of caring. But the one I want to focus on for the purposes of this essay—precisely because it's clearly, distinctly, and indeed self-evidently the case that it *cannot* be adequately modeled by any actual or really possible digital computing system or digital technology more generally—is the caring-based innate mental capacity that we all possess for what I call *principled authenticity*.

But before we get to principled authenticity, I'll need to sketch what I call *a Kantian-Frankfurtian conception* of our essentially embodied innate mental capacity for volition, in the specific mode of *willing*.

1. A Kantian-Frankfurtian Conception of the Will

Here is the pith-&-marrow of what Immanuel Kant[1] has to say about what he calls *practical freedom*:

> Freedom in the practical sense is the independence of the power of choice (*Willkür*) from necessitation by impulses of sensibility. For a power of choice is sensible insofar as it is pathologically affected (through moving-causes of sensibility); it is called an animal power of choice (*arbitrium brutum*) if it can be pathologically necessitated. The human power of choice is indeed an *arbitrium sensitivum*, yet not *brutum*, but *liberum*, because sensibility does not render its action necessary, but in the human being, there is a faculty of determining oneself from oneself, independently of necessitation by sensible impulses (*CPR* A534/B562, boldfacing in the original)

> Practical freedom can be proved through experience. For it is not merely that which stimulates the senses, i.e., immediate affects them, that determines human choice, but we always have a capacity to overcome impressions on our sensory faculty of desire by representations of that which is useful or injurious even in a more remote way; but these considerations about that which in regard to our whole condition is desirable, i.e., good and useful, depend on reason. Hence this also yields laws that are imperatives, i.e.,

[1] For convenience, I refer to Kant's works internally, that is, infratextually in parentheses. The citations include both an abbreviation of the English title and also the corresponding volume and page numbers in the standard "Akademie" edition of Kant's works: *Kants gesammelte Schriften*, edited by the Königlich Preussischen (now Deutschen) Akademie der Wissenschaften (Berlin: G. Reimer [now de Gruyter], 1902). For references to the first *Critique*, I follow the common practice of giving page numbers from the A (1781) and B (1787) German editions only. Because the Akademie edition contains only the B edition of the first *Critique*, I have also consulted the following German composite edition: *Kritik der reinen Vernunft*, ed. W. Weischedel, Immanuel Kant Werkausgabe III (Frankfurt: Suhrkamp, 1968). I generally follow the standard English translations of the German texts, but have occasionally modified them where appropriate. Here is a list of the abbreviations and English translations of the works cited.

CPR *Critique of Pure Reason*. Trans. P. Guyer and A. Wood. Cambridge: Cambridge Univ. Press, 1997.

CPrR *Critique of Practical Reason*. Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy.* Cambridge: Cambridge Univ. Press, 1996. Pp. 133-272.

GMM *Groundwork of the Metaphysics of Morals*. Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy.* Pp. 37-108.

IUH "Idea for a Universal History with a Cosmopolitan Aim." Trans. A. Wood. *In Immanuel Kant: Anthropology, History, and Education*. Cambridge: Cambridge Univ. Press, 2007. Pp. 107-120.

Rel *Religion within the Boundaries of Mere Reason*. Trans. A. Wood and G. di Giovanni. Cambridge: Cambridge Univ. Press, 1998.

> objective laws of freedom, and that say what ought to happen, even though it never does happen…. We thus cognize practical freedom through experience, as one of the natural causes, namely a causality of reason in the determination of the will (*CPR* A802-803/B830-831, boldfacing in the original)

And here is the pith-&-marrow of what Kant has to say about what he calls *transcendental freedom*:

> By freedom in the cosmological sense … I understand the faculty of beginning a state from itself (*von selbst*), the causality of which does not in turn stand under another cause determining it in time in accordance with the law of nature. Freedom in this signification is a pure transcendental idea, which, first, contains nothing borrowed from experience, and second, the object of which cannot be given determinately in any experience…. But since in such a way no absolute totality of [natural] conditions in causal relations is forthcoming, reason creates the idea of a spontaneity, which could start to act from itself, without needing to be preceded by any other cause that in turn determines it to action according to the law of causal connection (*CPR* A533/B561, boldfacing in the original).

Kant's theory of transcendental freedom is his metaphysics of free will. Transcendental freedom is how a rational human minded animal or human real person can, "from itself" (*von selbst*) (*CPR* A533/B561), be the spontaneous mental cause of certain natural events or processes. If I am that rational human minded animal or human real person, then insofar as I am transcendentally free, it follows that certain events or processes in physical nature are *up to me*—or to use Kant's own phrase, *in meiner Gewalt*, literally: "in my control" or "in my power" (*CPrR* 5: 94-95). So, otherwise put, transcendental freedom is deep freedom of the will, up-to-me-ness, or as it were (since it does not quite scan in grammatically correct German), *In-Meiner-Gewalt-Sein*. Now practical freedom presupposes transcendental freedom, and can be defined in a negative way as the independence of first-order volition, or the "power of choice" (*Willkür*), from necessitation by sensible impulses (*CPR* A533/B561). But practical freedom is also necessarily equivalent to what Kant calls "autonomy":

> the moral law expresses nothing other than the *autonomy* of pure practical reason, that is, [practical] freedom (*CPrR* 5: 33).

Practical freedom or autonomy is how a transcendentally free human real person can choose or do things by means of her subjective experience, or consciousness, of recognizing the Categorical Imperative or moral law as a desire-overriding, strictly universal, a priori, categorically normative, non-instrumental practical reason that has both motivating and

justifying force. In turn, the special fact of this subjective experience, or consciousness, of autonomous agency is what Kant calls the "fact of reason" (*Faktum der Vernunft*):

> The consciousness of this fundamental law [of pure practical reason, which says: so act that the maxim of your will could always hold at the same time as a principle of universal law giving] may be called a fact of reason, since one cannot ferret it out from antecedent data of reason, *such as the consciousness of freedom* (for this is not antecedently given), and since it forces itself upon us as a synthetic proposition a priori based on no pure or empirical intuition… In order to regard this law without any misinterpretation as given, one must note that it is not an empirical fact, but the sole fact of pure reason, which by it proclaims itself as originating law (*CPrR* 5: 31).

Otherwise put, the fusion of (i) transcendental freedom and (ii) practical freedom or autonomy, is what I'll call *free agency*.

As I interpret Kant's theory of free agency, this theory includes an early version of "the hierarchical-desire model of the will" later re-discovered by Harry Frankfurt (Frankfurt, 1988b, 1988e, 2004). According to Frankfurt's model, the rational human will is capable of having "effective first-order desires," which are desires for this or that, that do or will or would move us all the way to action, and also "second-order desires," which are desires about first-order desires, for example, the desire to desire this or that. Among our second-order desires, in turn, are special ones known as "second-order volitions," which self-consciously determine just which effective first-order desires are to be the ones that do indeed move us all the way to action. But in the specifically *Kantian* version of the Frankfurtian model, however, there are, among the second-order volitions, certain ones that are characteristic of what Kant and Frankfurt alike call *freedom of the will*, that operate by not only self-consciously determining just which effective first-order desires do indeed move us all the way to action (which expresses Frankfurtian freedom of the will), but also (and this is the specifically Kantian supplement) can either (i) de-rail an occurrent first-order desire which would otherwise have motivated the agent to action, or else (ii) newly-and-spontaneously generate an effective first-order desire that substitutes itself for an occurrent first-order desire which would otherwise have motivated the agent to action. On this Kantian-Frankfurtian hierarchical-desire model, then, *Willkür*, or the human power of choice/sensible will, is the faculty of effective first-order desires, or first-order volitions, and *Wille*, or practical reason (whether impure/instrumental or pure/non-instrumental), is the faculty of second-order volitions. The power of choice effectively desires ends or goals, and the satisfaction of desires produces pleasure or psychological happiness. Practical reason, in turn, recognizes the objective values of these ends or goals. When practical reason recognizes ends also as means for the production of happiness, it is instrumental. And when practical reason responds to ends for their own sake, it is non-instrumental. Practical

reason is therefore a spontaneous responsiveness of the agent to reasons, inhering in the structure of the rational human will itself, that guides the agent to intentional action.

Here, in turn, is Kant's basic theory of the constitution of the human will, which I also fully accept—or at least I fully accept it *as* I'm interpreting it, according to the Kantian-Frankfurtian hierarchical-desire model of the will I just sketched. The human will, or *faculty of desire* (*Begehrungsvermögen*), is our innate capacity for mobilizing and organizing our desires in order to motivate or move ourselves to choosing or doing, and in human persons the will is a rational human agent's power of wanting, intending, deliberating, deciding, and trying. In turn, the human will or the faculty of desire has two levels: (i) the lower or executive faculty of effective first-order desires, namely, first-order volitions, *the power of choice* (*Willkür*), and (ii) the higher or legislative faculty of second-order volitions, the will (*Wille*) proper, *or the faculty of practical reason.* So, the faculty of practical reason is a necessary proper part of the human will or faculty of desire. Moreover, the faculty of practical reason is the will in the proper or rational sense. In these ways, the faculty of practical reason is "the will proper" in two senses: it is an inherent part of the human will, and indeed the action-guiding core of the human will; and it also encapsulates the primary aim or end of the human will, which is to be a self-fulfilling rational human minded animal—that is, to be a creature that is principled and authentic, at least to some salient degree and extent. Otherwise put, according to my Kantian-Frankfurtian account, the human will or faculty of desire is a complex, integrated psychological structure that fuses desiderative animality and choice-or-action-guiding-and-controlling rationality,[1] and looks like this:

Human Will or Faculty of Desire (*Begehrungsvermögen)*

higher part = faculty of practical reason or will proper (*Wille)*

higher part = pure or non-instrumental reason

lower part = impure or instrumental reason

lower part = power of choice (*Willkür)*

In this Kantian-Frankfurtian model of the will, *Willkür*, or the power of choice, is an executive first-order volitional power of intentional causation by means of effective first-order desires. By contrast, *Wille*, or the will proper, is a higher-order volitional power of self-legislation that operates by means of recognizing either instrumental or non-instrumental reasons for the determination of choice. To act on the basis of *Willkür* is to move our animal bodies by means of our effective first-order desires. This can of course occur in a more or less Humean way by means of instrumental reasoning according to the hypothetical imperative. Since instrumental

[1] Tamar Schapiro aptly calls this a "middle road" approach to the nature of the will, i.e., one that neither identifies the will with desire nor over-intellectualizes the will (Schapiro, 2009).

reasoning is itself a form of self-legislation, it thereby involves what I'll call the "impure" *Wille*. Now the lower or executive faculty of desire, namely, the power of choice, is normally motivated or moved by objective ends that are picked out by our egoistic or self-interested, hedonistic, or consequentialist desires, and constitute the "matter" of our happiness, which is the pleasurable satisfaction of desires and the removal (or anyhow the control) of their painful frustration. I'll call this kind of happiness *shallow happiness* because it's essentially instrumental and psychological. Insofar as the faculty of practical reason is concerned with these ends, it is an "impure" and instrumental reason. This is the lower faculty of practical reason. But it is also possible for the faculty of practical reason to be pure and non-instrumental, and therefore to be moved not by the matter of our happiness, but rather solely by *the form of law-giving*, that is, by the structure of personhood or free agency itself, our essential nature as rational human minded animals, considered as an objective but purely formal end. This is the higher faculty of practical reason. The law that is given to themselves by human persons, that is, by free agents who are rational human minded animals, is the Categorical Imperative or moral law, hence higher willing of this type is positive freedom or realized autonomy.

3. The Rationality of the Heart: Principled Authenticity

The Highest or Supreme Good of Kant's ethics is having a *good will* in Kant's sense (*CPrR* 5: 110-111; *GMM* 4: 393), that is, occurrently partially or completely realizing autonomy.[1] For my own part, I've interpreted and updated the Kantian notion of the Highest Good as *a unified system of transcendental values* that is centered on *human dignity*, i.e., the absolute, non-denumerably infinite, intrinsic, and objective value or worth (*Würde*) of human real persons (Hanna, 2023a, 20233b, 2023c). *The Realm of Ends*, also called "the ethical community" (*Rel* 6: 94), is the total moral community of rational human minded animals or human real persons, each of whom sufficiently respects one another and themselves as creatures with dignity (absolute, non-denumerably infinite, intrinsic, and objective value or worth), and also considers all the others and themselves equally in relation to the Categorical Imperative/moral law; and, finally, each possesses a good will. The "sole and complete good," (*GMM* 4: 396)—that is, the

[1] Confusingly, in the *Critique of Practical Reason* (*CPrR* 5: 110-111), Kant uses the term "the highest good" to mean the same as what he had called "the sole and complete good" in the *Groundwork*, namely the moral totality consisting of human happiness proportioned to moral virtue (*GMM* 4: 396). The notion of this moral totality, in turn, is the same as the notion captured by the third Postulate of Pure Practical Reason, the moral belief in God's existence as a regulative Idea, since only an all-powerful and all-good God could create such a moral totality (*CPrR* 5: 124-132). By contrast, in the *Groundwork*, "the highest good" means *a good will* (*GMM* 4: 393-394), which is the same as what he calls "the supreme good" in the second *Critique* (*CPrR* 5: 111). To be sure, all of these slightly different notions are essentially connected, since the supreme good (namely, a good will) is the formal essence and immanent structure of the sole and complete good (that is, the moral totality consisting of human happiness proportioned to moral virtue) that only an all-powerful and all-good God could create.

best life for any rational human minded animal or human real person—is a life of deep individual happiness and also deep communal or social happiness, that's "deep" precisely insofar as as it's intrinsically controlled and structured by a good will in the Kantian sense, and not merely "shallow" or essentially instrumental, psychological happiness.

To act on the basis of our "pure" *Wille*, our own capacity for pure practical reason, however, is to constrain and determine our *Willkür*, our own power of choice, by recognizing the Categorical Imperative. Insofar as it is recognized by us, the Categorical Imperative provides a desire-overriding, strictly universal, a priori, non-instrumental reason for action. That recognition, in turn, causally triggers an innate higher-order emotional disposition in us (also known as the capacity for *Achtung* or respect) to generate the consciously experienced desire to be moved by morally appropriate and non-egoistic or non-self-interested, non-hedonistic, non-consequentialist effective first-order desires:

> The direct determination of the will by the law, and the awareness of that determination, is called "respect," so we should see respect as the *effect* of the law on a person rather than as what *produces* the law. Actually, respect is the thought of something of such worth that it breaches my self-love…. Any moral so-called *interest* consists solely in *respect* for the [moral] law (*GMM* 4: 402 n.)

Respect, in turn, is essentially a capacity and (when the capacity is triggered) an operation of the rationally-responsive *human heart*:

> The pure thought of duty and in general of the moral law, mixed with no foreign addition of empirical inducements, has by way of reason alone (which with this first becomes aware that it can of itself be practical) an influence on the human heart so much more powerful than other incentives, which may be summoned from the empirical field, that reason, in the consciousness of its dignity, despises the latter and can gradually become their master (*GMM* 4: 410-411, underlining added).

> [T]hat a human being should become not merely legally good, but morally good (pleasing to God) i.e. virtuous according to the intelligible character [of virtue] … and thus in need of no other incentive to recognize a duty except the representation of duty itself—that, so long as the foundation of the maxims of the human being remains impure, cannot be effected through gradual reform but must rather be effected through a revolution in the disposition of the human being…. And so, a "new man" can come about only through a kind of rebirth, as it were a new creation (John, 3:5; compare with Genesis, 1:237) and a change of heart.

> But if a human being is corrupt in the very ground of his maxims, how can he possibly bring about this revolution by his own forces and become a good human being on his own? Yet duty commands that he be good, and duty commands nothing but what we can do. The only way to reconcile this is by saying that a revolution is necessary in the mode of thought but a gradual reformation in the mode of sense (which places obstacles in the
>
> way of the former), and [that both] must therefore be possible also to the human being. That is: If by a single and unalterable decision a human being reverses the supreme ground of his maxims by which he was an evil human being (and thereby puts on a "new man"), he is to this extent, by principle and attitude of mind, a subject receptive to the good; but he is a good human being only in incessant laboring and becoming, i.e., he can
>
> hope—in view of the purity of the principle which he has adopted as the supreme maxim of his power of choice, and in view of the stability of this principle—to find himself upon the good (though narrow) path of constant progress from bad to better. For him who penetrates to the intelligible ground of the heart (the ground of all the maxims of the power of choice), for him to whom this endless progress is a unity, i.e. for God, this is the
>
> same as actually being a good human being (pleasing to him); and to this extent the change can be considered a revolution [of the heart]. For the judgment of human beings, however, who can assess themselves and the strength of their maxims only by the upper hand they gain over the senses in time, the change is to be regarded only as an ever-continuing striving for the better, hence as a gradual reformation of the propensity to evil, of the perverted attitude of mind. (*Rel* 6: 47-48, underlining added)
>
> We betray a culpable degree of moral unbelief if we do not grant sufficient authority to duty's precepts, as originally inscribed in the heart by reason. (*Rel* 6: 84, underlining added)
>
> Moral faith (*Glaube*) must be a free faith, founded on pure dispositions of the heart (*fides ingenua*). (*Rel* 6: 115, underlining added)
>
> The highest goal of the moral perfection of finite creatures, never completely attainable by human beings, is … the love of the [moral] law. (*Rel* 6: 145, underlining added)

In this way, to choose or act on the basis of pure *Wille* is to do the good and right thing, as determined by our own pure practical reason, via the unique, wholly heartfelt motivational

influence of the innate dispositional higher-order emotion of respect on our effective first-order desires or choices, no matter what the external and psychological antecedents, no matter how much pain I might suffer by doing the right thing, and no matter what the consequences.

The crucial factor in this account is Kant's truly important idea, also fully endorsed by me, that there exists an innate emotional disposition in all rational human minded animal agents to experience spontaneously generated wholehearted higher-order desires to be moved by non-egoistic or non-self-interested, non-hedonistic, non-consequentialist effective first-order desires or choices. I call this special wholehearted higher-order desire *the desire to achieve self-transcendence* because it is a person-unifying, life-changing desire to achieve a radical volitional distancing with respect to our own egoistic or self-interested, hedonistic, or consequentialist first-order desires, and thus to be able to overcome the almost irresistible centripetal forces of The Dear Self and The Bottom Line. Now non-egoistic or non-self-interested, non-hedonistic, non-consequentialist first-order desires take the following general form: I want (not-) X, no matter how much egoistic or hedonistic unhappiness and pain I may experience in getting (not-) *X*, and no matter what the consequences.

So, correspondingly, the higher-order desire to achieve self-transcendence takes the following general form: I want (not) to want (not-) *X*, no matter how much egoistic or hedonistic unhappiness and pain I may experience in getting (not-) *X*, and no matter what the consequences.

Here, in turn, are two key points about the desire to achieve self-transcendence.

First, postulating the desire to achieve self-transcendence as the motivational ground of chooing or acting for the sake of the Categorical Imperative or moral law is *not* "emotional extremism" according to any specifically negative philosophical or moral connotation of that phrase.[1] On the contrary, it is perfectly rational and human—indeed a proper part of the authentic self-fulfillment of practical human rationality—to be prepared to go to the wall for the things that really and truly matter most to you. It is not that you specifically *want* to go to the wall or that you are specifically *trying* to go to the wall. Indeed, it is a necessary part of the desire to achieve self-transcendence when it is specifically triggered by recognition of the Categorical Imperative, that *were there any other relevant options short of that*, which also sustained the highest or supreme good, and also brought about good consequences for others and for oneself, *then you would go for that instead*. It is just that the highest or supreme good is *worth more* than any actual or possible set of good consequences alone. Therefore, good consequences do not necessarily determine the highest or supreme good.

It's important to note that trying to bring about good consequences for others and oneself, other things being equal, is a strict moral obligation, and hence good consequences for others

[1] See, e.g., Derek Parfit: "Kant is sometimes thought of as a cold, dry, rationalist. But he is really an emotional extremist" (Parfit, 2011, xliv).

and oneself constitute a proper part of the highest or supreme good (= a good will = a life of principled authenticity, achieved at least partially or to some degree = rational human minded animal agency that sufficiently respects everyone's human dignity), according to the sketch of Kant's theory of free agency I'm developing here, and also work out in full detail and defend in my book, *Kantian Ethics and Human Existence* (Hanna, 2018b). The crucial thing to note here and now, however, is just that any attempt to substitute *that proper part* of the highest or supreme Good (= good consequences for others and oneself) for *the whole* of the highest or supreme Good (= a good will = a life of principled authenticity, achieved at least partially or to some degree = rational human minded animal agency that sufficiently respects everyone's human dignity) is fallacious.[1] The whole highest or supreme good, taken by itself, is worth nondenumerably infinitely more than any of its proper parts, and has its value whatever the consequences.

The second key point is that *sometimes* the desire to achieve self-transcendence is near-Satanically evil, and thus undertaken under the Guise of the Bad, as in a case in which someone cold-bloodedly tries to murder or seriously harm someone else, *purely for the sake of violating (respect for) their dignity*, even though he knows that he is bound to be shot by the police. In such a case, the intrinsic value, or objective end, that triggers the higher-order desire to achieve self-transcendence is the fact that the assailant chooses and acts on the morally disvalued and wrong goal and reason, no matter how much pain he may experience in bringing it about, and no matter what the consequences.

As I say, this is a clear and distinct case of near-Satanic evil. Near-Satanic evil minimally implies our ability to act with transcendental freedom of the will, namely, agentive sourcehood or up-to-me-ness, but also in a monstrously self-loving/narcissistic and self-inflating/selfish way, and wrongly, hence without practical freedom of the will, a good will, or occurrently realized autonomy. Of course it must also be added that both the capacity for and also the realization of transcendental freedom entail our possession of the capacity for practical freedom, since all of these can occur only in a rational human minded animal agent, and amongst the rational minded animal agents, only (as far as we know, leaving aside possible rational alien or Earthly nonhuman rational minded animal agents) in a rational human minded animal agent (*CPR* A533-534/B561-562). But near-Satanic evil *also* implies our ability to act freely on the basis of innately-specified and spontaneously generated, highly maleficent, but also non-egoistic, non-hedonistic, or non-consequentialist *first-order* desires. It is possible wholeheartedly to want a thing that flagrantly violates human dignity, no matter how much first-order egoistic and hedonistic unhappiness and pain you experience in getting it, and no matter what harmful things to you or anyone else as a consequence of your actions. So, you

[1] This line of thinking is closely related to G.E. Moore's arguments in *Principia Ethica* against The Naturalistic Fallacy, and in support of the primitiveness and irreducibility of The Good (see Moore, 1903; Hanna, 2022).

want that evil thing for its own sake, literally *for the hell of it*. In this way, just like David Hume, Kant does not regard it as contrary to reason for someone to prefer the destruction of the world, including his own self-destruction, to the scratching of his finger (Hume, 1978: book II, part III, section iii, 416).

In order to make this rational human emotional possibility of near-Satanic evil even more concrete, we can think not only of Hume's monstrously narcissistic/self-loving and selfish/self-inflating person, but also of John Milton's Satan, a literary character that Kant in all likelihood would have known well (Budick, 2010), of Iago in Shakespeare's *Othello*; of the nihilist Peter Verhkovensky in Dostoevsky's *The Devils*; of the post-modern punk thug Alex in Anthony Burgess's/Stanley Kubrick's *Clockwork Orange*; of the brilliant serial-killer/cannibal Hannibal Lecter in Jonathan Demme's *Silence of the Lambs*; of the ultra-cold-blooded hitman Anton Chigurh in Cormack McCarthy's novel/the Coen Brothers' movie *No Country for Old Men*; of the real-life totalitarian mass-murderers, Hitler and Stalin; of the real-life perpetrators of the Columbine High School Massacre in 1999, Eric Harris and Dylan Klebold; of the homegrown US terrorist Timothy McVeigh, federally executed in 2001 for the Oklahoma City Bombing in 1995; and, sadly, of many, many other moral monsters, too numerous to mention.

Such evil, although it is monstrously egoistic, and near-Satanic, is *nonideally rational*. Only a nonideally rational, even though "human, all too human," animal could ever have such a self-transcending desire. Indeed, on Kant's account of the nature of desire, no desires had by rational human minded animals could ever be essentially *irrational* or *arational*, since the function of a desire in a rational human minded animals is just to move themselves to action in the service of attaining rationally-recognized objectively intrinsically valuable instrumental or non-instrumental ends—whether these are material ends, in the case of empirical desires based on pleasure and pain, or formal ends, in the case of moral emotion of respect (*CPrR* 5: 21-28). But some non-egoistic, non-hedonistic, non-consequentialist desires are more rational than others, and some are monstrously egoistic. So according to Kant, it would be monstrously egoistic for me to prefer the destruction of the world (including my own self-destruction) to the scratching of my finger, precisely because this would be a radical violation of the Formula of Humanity as an End in Itself version of the Categorical Imperative:

> So, act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means. (*GMM* 4: 429)

I would thereby be considering and treating everyone in the world, including myself, as mere things and mere means to my own ends, and worth less than my momentary mild pain, for its own categorically wicked sake and no matter what.

"For its own categorically wicked sake, and no matter what." Therefore, in an apparent paradox, near-Satanic evil is *the exact reversed mirror-image or enantiomorph of acting with*

a good will, for good ends, without self-interest, non-hedonically, for its own categorically good sake, whatever the consequences—the "flip side" of a Kantian moral saint. Correspondingly, the other seemingly paradoxical thing about near-Satanic evil, as per Kant's profound idea of a "revolution of the heart" or "revolution of the will" that we looked at above (*Rel* 6: 47-48) and as per Dostoevsky's novels, given the transcendentally free, low-bar rational, double-edged potentiality of the desire for self-transcendence, is that it only takes *one* life-changing, existential Kantian act of choice to go over completely from Great Sinner to Saint.

Now, in your first-person moral imagination, starting out as a Great Sinner or as a banal sinner, as a Hitler/Stalin or as an (Arendtian) Eichmann, or even just as an everyday non-evil, morally bad person, *revolutionize your will*, and if you are evil, then *flip the flip side*. Therefore, the desire for self-transcendence, whenever it results from recognition of the Categorical Imperative/moral law, thereby motivating choice or action for the sake of the Categorical Imperative/moral law, is inherently morally good and right. That inherently morally good and right self-transcendence only *rarely* happens in human affairs is fully acknowledged by Kant: "from the crooked timber of humanity, nothing straight can ever be made" (*IUH* 8: 23). But on the other hand, Kant firmly believed, even given the inherent crookedness of our timber, our sheer bent humanity, it *is* really possible.

In support of this, he provides a famous thought-experiment of a very lustful person who (sharply unlike the near-Satanic and monstrously egoistic Humean person, who prefers the destruction of the world, including his own self-destruction, to the scratching of his finger) would never in fact gratify his lust and thereby commit a crime for any instrumental reason, if at the moment of committing his crime he were presented with the gallows from which he would be instantly strung up as punishment. But this very same very lustful person nevertheless regards it as fully possible for him to lay down his own life on the very same gallows by refusing to give false testimony against an honorable man, even though he were commanded to do on pain of death by a tyrannical prince, and thus the very lustful person conceives it to be really possible for him to choose and act on the basis of a moral non-instrumental reason (*CPrR* 5: 30).

But how is this radical "life-change," "reversal," or "revolution" of the human heart, that is, of the power of choice/sensible will, really possible, for a very lustful or even near-satanically evil person? The quick Kantian answer is that it *can* happen because it *ought* to happen, *and* because we all actually *do* innately have the power (for transcendental freedom, practical freedom, and autonomy, i.e., free agency) to make it happen. The longer Kantian answer is that it results from our spontaneous power of choice together with the intense personal struggle to reconfigure the basic hierarchical-desire-structure of one's own embodied will so as to achieve increasingly greater degrees of moral authenticity and self-transcendence, thereby overcoming egoism. How few of us ever manage to do this! And even for those few who do, how

infrequently it ever happens! But in any case, frankly, I also think it actually *does* happen sometimes, and *more often* than you might think.

To be sure, publicly-acknowledged examples of moral heroism, sainthood, or radical goodness and rightness in the Kantian sense are not, to put it mildly, as thick on the ground as the relentlessly repetitive, Wheel of Ixion-like, trillionfold examples of near-Satanic or banal evil, and moral badness. Moreover, many or even most examples of moral heroism, sainthood, or radical goodness and rightness in the Kantian sense consist in all-too-tempting evil things *not* chosen and *not* done. Consider this imaginary headline Woman Resists Terrible Temptation to Do Evil; Intensely but Quietly Suffers for It; Experiences Moral Self-Fulfillment; Then Goes About Her Daily Life as Usual.

Such stories rarely make it into *USA Today* or onto Fox News, nor even into *The New York Times* or *WaPo*. I mean, of course, that they never do and they never will. And the representation of radical goodness and rightness in art, sadly, also tends to be moralistic and prissy.

But in order to make the rational human emotional possibility of radical goodness and rightness more concrete, we can think here of Socrates as represented by Plato in the *Apology*, the *Crito*, and the *Phaedo*; of the absurd "Knight of the Sorrowful Countenance," Don Quixote, in Cervantes's *Don Quixote*; of Kierkegaard's "Knight of Faith" in *Fear and Trembling*; of the "Idiot" Prince Myshkin in Dostoevsky's *The Idiot*; of Renée Falconetti's brilliant portrayal of Joan of Arc in Carl Theodor Dreyer's *Passion of Joan of Arc*; of Takashi Shimura's equally brilliant portrayal of the dying civil servant Kanji Watanabe in Kurosawa's *Ikiru*; and also of the real-life, and therefore "human, all too human," but still genuine moral heroes Frederick Douglass, Abraham Lincoln, Mahatma Ghandi, Martin Luther King Jr., and Mother Teresa. And leaving aside these famous real-world cases, I also strongly believe that there are some people you actually know, or at least have met, or have heard about, who, in their own (perhaps) less consistent and (certainly) less conspicuous ways, are relevantly like this. And if *they* can be like that, and choose and act like that, then *you* can too.

So according to Kant, the real possibility of radical human goodness and rightness via a radical life-change, reversal, or revolution of the human heart is just an actual fact, although of course a unique sort of fact, namely a non-empirical fact about the rational human minded animal heart. More precisely, however, this unique non-empirical fact is the fact that our subjective experience or consciousness of recognizing the Categorical Imperative/moral law triggers our innate higher-order emotional disposition for feeling respect, that is, our love for the moral law that's inscribed in the human heart, and then the moral feeling of respect spontaneously generates the wholehearted higher-order desire to achieve morally right self-transcendence: *straight from the heart*. So, it is not only a *non-empirical* fact, but also an inherently *affective* fact. This subjective experience, or consciousness, of our direct recognition of the Categorical Imperative/moral law, together with its higher-order, causal-generative

effects, is nothing more and nothing less than The Fact of Reason, and the text in which Kant introduces it bears repeating:

> The consciousness of this fundamental law [of pure practical reason, which says: so act that the maxim of your will could always hold at the same time as a principle of universal law giving] may be called a fact of reason, since one cannot ferret it out from antecedent data of reason, such as the consciousness of freedom (for this is not antecedently given), and since it forces itself upon us as a synthetic proposition a priori based on no pure or empirical intuition… In order to regard this law without any misinterpretation as given, one must note that it is not an empirical fact, but the sole fact of pure reason, which by it proclaims itself as originating law. (*CPrR* 5: 31, underlining added; see also 5: 6, 42-43, 47-48, 55-57, 91-94, and 104-108)

Strictly speaking, Kant often refers to this as *a* fact of reason—although of course in this text he does refer to it as "the sole fact of pure reason"—but I will continue to refer to it as *The* Fact of Reason, simply in order to distinguish it from other actual or possible facts of reason that may not be strictly speaking *moral* facts of reason, such as the regulative use of ideas of pure reason in natural scientific inquiry (*CPR* A642-652/B670-680) and the closely-related regulative use of ideas of pure reason in biology and teleological contexts more generally (*CPJ* 5: 167-198).

It should be especially noted that, as I'm construing it, The Fact of Reason is *perfectly consistent with* this famous text in the *Groundwork*:

> In fact, it is absolutely impossible by means of experience to make out with complete certainty a single case in which the maxim of an action otherwise in conformity with duty rested simply on moral grounds and on the representation of one's duty (*GMM* 4:407).

This famous text from the *Groundwork* has always—or almost always—been taken to imply a strong and specifically Kantian version of moral epistemic skepticism, by saying that we cannot ever know with certainty whether we are acting or have acted *for the sake of* the Categorical Imperative, that is, *from duty*, and not merely *in conformity with duty*, but actually from egoistic or hedonistic or utilitarian motives. But my reading of this famous text is sharply different. I think that the crucial sub-phrase in Kant's key phrase, "it is absolutely impossible by means of experience to make out with complete certainty a single case…," is "*by means of experience*." More precisely, then, I think that it is perfectly consistent with what Kant has actually written at *Groundwork* 4: 407, to claim that, *yes*, *empirical knowledge*, in Kant's technical senses of "empirical" and "knowledge" (*Wissen*) of whether we are acting or have acted from duty or not, is impossible. Empirical knowledge for Kant necessarily involves empirical concepts and

empirical intuitions, and has its meaning, justification, and truth necessarily determined, at least in part, by sensory, contingent facts about the natural world and ourselves. Moreover, at most, empirical knowledge can have *empirical certainty*, which is when a subject "takes something [namely, a judgment or proposition] to be true" (*Fürwahrhalten*), in such a way that this taking-to-be-true has a kind of indubitability or self-evidence which is both subjectively sufficient, which makes it "conviction" (*Überzeugung*) and also intersubjectively or objectively sufficient (*CPR* A820-822/B848-850), which entails that the judgment or proposition is also actually true. Examples of empirical certainty would be cases of ordinary perceptual judgments under highly favorable contextual conditions, such as G.E. Moore's famous anti-skeptical thesis, "This is my right hand and this is my left hand." I think that Kant is absolutely correct that we cannot ever have knowledge *in this sense* of whether we are acting or have acted from duty or not.

Nevertheless, even while conceding that empirical knowledge of the morality of our own actions is impossible, we can *also* consistently and also justifiably claim on Kantian grounds that we still have the capacity for *veridical, direct, occurrent awareness* of our choosing and acting from from duty, by means of *non-empirical moral self-consciousness.* In other words, even though we cannot have empirical knowledge of whether we are acting or have ever acted from duty or not, we can still have *non-empirical certainty* about this. Correspondingly, I also think that The Fact of Reason is precisely this veridical, direct, occurrent, non-empirical moral self-consciousness of our acting from duty. Therefore, given The Fact of Reason, then we can indeed have veridical, direct, occurrent moral self-consciousness, with non-empirical certainty, when (and only when) we are acting for the sake of the Categorical Imperative and from duty. Otherwise put, acting from duty *is partially constituted by its own self-validating phenomenology*.

In any case, what is most crucial here is to note, again, is that The Fact of Reason is not a conceptual, propositional, inferential, or more broadly speaking *intellectual* fact, but instead an inherently heartfelt or *affective* fact about how the higher-order moral feeling of respect, that is, love for the moral law inscribed in the human heart, operates on the hierarchical-desire-structure of the human will. So The Fact of Reason is really *The Affect of Reason*. All rational facts are either absolutely (that is, without requiring empirical inputs) or relatively (that is, requiring empirical inputs) spontaneously active; and the Fact or Affect of Reason is absolutely spontaneously active, insofar as it is absolutely spontaneously *passionate*. In other words, The Fact or Affect of Reason is a rational act of the human heart, not merely a rational act of the head. In this respect, Kant's view is strikingly like that of Pascal, who very famously wrote that

> [t]he heart has reasons of its own, that reason knows nothing about. (Pascal, 1955: 154, translation modified slightly)

As I read that profound Pascalian thought from a Kantian point of view, it's saying that the human heart has (RH, on behalf of Kant: pure practical or moral) reasons of its own that (RH, on behalf of Kant: theoretical or scientific) reason knows nothing about.

For example, someone raises her arm and shrieks in order to stop a street crime, or perhaps she becomes a whistleblower in a corporate, governmental, or scientific organization (see, e.g. Sarasohn, 1993[1]), just because she feels in her essentially embodied heart that it's the morally right thing to do, even though she thereby risks her own life (in the case of stopping the street crime), or even though she risks losing her job and all her co-worker friends, and perhaps also ruining her career and her marriage (in the case of the corporate, governmental, or scientific whistleblower), and even though she desperately wants to avoid "being involved." It seems clear that given these background conditions, only a second-order volition driven by the innate affective capacity for respect, that is, a radical kind of love, could motivate such acts. Therefore, she is doing her duty. According to Kant, "duty is the necessity of an action [done] from respect for the moral law" (*GMM* 4: 400). This says, as I'm understanding it, that duty is the obligation that's binding on any act which is such that *only* the feeling of respect for persons and the Categorical Imperative/moral law innately inscribed in the human heart will suffice to move us, no matter what our first-order desires might happen to be, and no matter what the consequences.

In turn, there seem to be two importantly distinct ways in which the feeling of respect can move us by way of the second-order volition consisting of the more-or-less wholehearted desire to achieve a morally principled self-transcendence.

According to the first way, the higher-order wholehearted desire to achieve morally principled self-transcendence can take a particular online egoistic, hedonistic, or merely consequentialistic would-be effective first-order desire offline, and substitute a morally appropriate pre-existing or latent (but as yet non-effective) non-egoistic, non-hedonistic, non-consequentialistic (hence non-instrumental) first-order desire in its place, so that it becomes the effective one. To borrow Kant's example, the very lustful person can take his intense online first-order desire to avoid being hanged offline, and then substitute a pre-existing or latent first-order desire to avoid bearing false witness against an honorable man, so that this latter desire now becomes his first-order volition.

And according to the second way, assuming a complete set of selfish, egoistic, hedonistic, or consequentialist (hence instrumental) online first-order desires, together with another complete set of non-egoistic, non-hedonistic, or non-consequentialist (hence non-instrumental) first-order desires, from which none has yet emerged as a would-be effective first-order desire, the higher-order wholehearted desire to achieve principled self-transcendence can spontaneously re-organize the emotional constituents of that state so as to produce a new online non-egoistic, non-hedonistic, or non-consequentialist (hence non-instrumental) effective first-

[1] This is an account of the notorious "David Baltimore case."

order desire that is also morally appropriate. To borrow another of Kant's examples, a person who is by nature somewhat cold and unsympathetic towards other people, and furthermore has many troubles of his own, can nevertheless wholeheartedly generate a new effective first-order desire to be kind to someone else (*GMM* 4: 398-399). In turn, this sort of emotionally and practically generative absolute spontaneity is strictly analogous to the intellectually and theoretically generative absolute spontaneity that yields pure a priori knowledge (Hanna, 2006: ch. 7).

Now, again, I'm not saying, nor is Kant saying, that this is an easy thing to do, nor am I saying, nor is Kant saying, that it happens very frequently in the course of everyday, personal or public life. But in fact, it actually happens a lot more often than you might think, as Rebecca Solnit's brilliant study of spontaneously-formed, altruistic "disaster-communities," *A Paradise Built in Hell*, and Larissa MacFarquhar's equally brilliant study of habitually or spontaneously altruistic people, *Strangers Drowning*, both compellingly show (Solnit, 2009; MacFarquhar, 2016). In any case, I think it's clearly really possible, and, just as clearly, sometimes actual. Indeed, as I indicated above, I also think that to some salient degree, *we have all either directly experienced this disposition or propensity in ourselves or else clearly recognized it in others*. At the very least, we have all recognized that we are fundamentally *capable* of it, because we have all recognized, no matter how fleetingly or darkly, *that a human real person can and should change her life for the better*. For example, consider one of the best-known lines in modern poetry, the last line of Rilke's "Archaic Torso of Apollo": *Du mußt dein Leben ändern*— "You must change your life" (Rilke, 1989, 61). Rilke's *poetic* imperative is clearly a Kantian *moral* imperative. Now, insofar as we can read Rilke's poem, and understand this line, then we morally *imagine* changing our own lives radically for the better, and rationally *hope* that we can.

We should also *not* assume, however, that the deeply motivational, desire-overriding, innate emotional disposition for feeling respect for persons and the moral law, i.. e., the wholehearted desire to achieve morally principled self-transcendence, at least to some salient degree or extent, when triggered into action, will always have *the same phenomenology*. It may manifest itself as intense feelings of depression or gloominess (for example, Lincoln); of guilt; of sympathy; of empathy; of ecstatic suffering (for example, Falconetti's Joan of Arc); of intense outwardly-directed anger; or even of self-loathing (for example, Prince Myshkin). As Kant points out, since it "breaches my self-love" or narcissism, and also "humiliates," "strikes down," and "weakens" my self-inflation or selfishness, the subjective experience of respect or the desire for morally principled self-transcendence is often extremely unpleasant.

Of course, other things being equal, it isn't terribly enjoyable, indeed "it's not my idea of a good time," to thwart one's own powerful egoistic, hedonistic, or consequentialist first-order desires. Freudians would call it *repression*, and this also raises a corresponding philosophical

worry: If having a good will in the Kantian sense often involves repression, hence "what's not my idea of a good time," then how can it still be *good*? In reply to the Freudians, Kant could say:

> "Yes, I agree completely that it is repression, and also that repression, other things being equal, is "what's not my idea of a good time": not a shallowly happy experience. But precisely because we are crooked timbers and radically evil, a certain amount of serious repression is just the psychic cost of moral goodness, rightness, and virtue. The vitally important point is *that it is more-or-less wholehearted*, not *that it is normally either pleasant or self-gratifying*. Acting with a good will involves psychic coherence and self-sufficiency, not necessarily shallow happiness."

This vitally important point requires more elaboration. There's repression, and then there's repression. Certainly, much repression is intensely unpleasant, morally and personally pointless, and even positively harmful. Think, for example, of all the thoroughly messed-up central characters in Hitchcock's most brilliant films from the notoriously repressed 1950s: *Strangers on a Train*, *Vertigo*, *Rear Window*, and especially (at the very end of the decade) *Psycho*. But a transcendentally free and practically free or autonomous rational human minded animal—that is, a conscious, self-conscious, and self-reflective human agent, capable of theoretical and logical a priori cognition, who also has the innate emotional and practical capacity for being motivated or moved by respect for persons and the moral law, or by the wholehearted desire to achieve morally principled self-transcendence—may *sometimes* be, but does not ever *have to* be, helplessly handcuffed, manipulated, overwhelmed, twisted, or violated by her own desires. This is because the innate emotional disposition for feeling respect for persons and the moral law, or the desire to achieve principled self-transcendence, essentially affectively expresses her deepest self.

Think again here of Falconetti's Joan of Arc, and of Mother Teresa, and of all those thousands or even millions of obscures, inconspicuous, unsung, perhaps only part-time, everyday moral saints that everyone actually knows, has met, or at least has heard about. If a rational human minded animal agent is ever truly motivated or moved by respect for persons and the moral law, or by the wholehearted desire to achieve principled self-transcendence, even if it requires a terrible struggle to achieve it, then ultimately, she has the will that she most deeply wants. She has thereby realized the capacity for rational emotional control of her own conscious, affective, and practical life. The internal constitution of the person she is and the person she will become are then both ultimately *up to her*: they flow from her, as the agential source. She is therefore both transcendentally (aka deeply) free and also practically free, hence partially or completely occurrently autonomous.

Kant very aptly calls the subjective experience, or consciousness, of this special sort of moral self-control and agentive sourcehood, "self-fulfillment" or *Selbstzufriedenheit*:[1]

> Have we not, however, a word that does not denote enjoyment, as the word happiness does, but that nevertheless indicates a satisfaction with one's existence, an analogue of happiness that must necessarily accompany consciousness of virtue? Yes! This word is *self-fulfillment*, which in its strict meaning always designates only a negative satisfaction with one's existence, in which one is conscious of needing nothing. Freedom, and the consciousness of freedom as an ability to follow the moral law with an unyielding disposition, is *independence from the inclinations*, at least as motives determining (if not as *affecting*) our desire, and so far, as I am conscious of this freedom in following my moral maxims, it is the sole source of an unchangeable fulfillment, necessarily combined with it (*CPrR* 5: 117)

Such a state of rational, volitional self-fulfillment is a higher, and indeed a *higher-order*, kind of happiness that is analogous in certain respects to first-order shallow happiness, in that it results from the satisfaction of desires, but also sharply different in that it consists in the satisfaction of a *special higher-order desire*—the desire for self-transcendence—not in the satisfaction of first-order desires. Moreover, it is essentially *deeper* than first-order shallow happiness. As "a negative satisfaction with one's existence (*ein negatives Wohlgefallen seiner Existenz*), in which one is conscious of needing nothing," it is the emotional state of rational, volitional coherence and self-sufficiency in a contingent, complex, and thoroughly nonideal actual world. Or as Kant puts it in a text that shows him at his affectively intelligent best:

> *Considered in themselves* natural inclinations are *good*, i.e., not reprehensible, and to want to extirpate them would not only be futile but harmful and blameworthy as well; we must rather constrain them, so that they will not wear each other out but will instead be harmonized into a whole called happiness (*Rel* 6: 58, underlining added).

Higher, higher-order happiness as *Selbstzufriedenheit* is Kant's anticipation of what the Existentialists later called "authenticity," or *Eigentlichkeit*, including what Kierkegaard called "purity of heart"—

[1] *Selbstzufriedenheit* could also be translated as "self-contentment," "self-gratification," or "self-satisfaction." But in English these have unfortunate and misleading connotations of passive, sanctimonious, smugness or merely egoistic enjoyment. "Self-fulfillment" avoids these, and also captures the inherently teleogical character of Kant's idea, because it fully includes the notions of self-determination and self-realization.

> *Purity of Heart Is to Will One Thing* …. [T]he person who in truth wills only one thing *can will only the good*, and the person who wills only one thing when he wills the good *can will only the good in truth* (Kierkegaard, 2000, 271).

also, of what Frankfurt calls the "decisive identification" of second-order volitions with effective first-order desires or first-order volitions (Frankfurt, 1988b, 21), and, when it is dynamically spread out over time, "wholeheartedness"—although, of course, in the present context with an inherent and specifically Kantian orientation towards respect for real persons and the Categorical Imperative/moral law. As I've said, I call this crowning achievement of free agency *principled authenticity*. Whatever we call it, along with Kant I do think it is a variety of free will most definitely worth having: indeed, Kant in my opinion altogether rightly thinks that it is the variety of free will *most* worth having.

Human practical reason is our vital, relatively or absolutely spontaneous, capacity to exercise the power of choice for the sake of instrumental or non-instrumental principles; and *Selbstzufriedenheit* is the subjective experience of partially or completely wholeheartedly realizing this capacity. Then, *in*authenticity in the specifically Kantian sense, or the moral contrary of *Selbstzufriedenheit*, is just to comport yourself "heartlessly," or as if you were nothing but a puppet, natural automaton, or machine—as if you were nothing but a *Dalek* in the famously long-running British TV science fiction series *Dr Who*—wholly determined by natural causal laws, and neither alive nor practically free. Or in other words, inauthenticity in the Kantian sense is the self-automating denial of your own capacity for practical freedom:

> [I]f the freedom of our will was nothing else than [an *automaton spirituale* when it is impelled by representations] … it would in essence be no better than the freedom of a turnspit, which when once wound up also carries its motions from itself. (*CPrR* 5: 97, underlining added)

The doctrine of *Selbstzufriedenheit*, in turn, highlights the basic way in which Kant's theory of free agency transcends Hume's theory of practical agency. For Kant, practical reason is *not* "the slave of the passions" (Hume, 1978: book II, part III, section iii, 415). But this does not imply that for Kant, practical reason is *not* intrinsically connected to our desires, drives, emotions, and feelings, and thus intrinsically connected to our passions. On the contrary, according to my account of Kantian free agency, practical reason *is* intrinsically connected to our passions, and indeed intrinsically connected to our deepest and most self-expressive passions: the ones whose ends we would be prepared to go to the wall for. These passions are *the vital engines* of pure practical reason, and practical reason is *the non-mechanical, relatively or absolutely spontaneous, first-order conscious and also self-conscious structural guidance-and-control system for these engines*. Via our faculty for practical reason, we consciously and self-

consciously recognize the relative or absolute objective intrinsic values of ends; and at the very same time and in the same respect, our desires, drives, emotions, and feelings subjectively and more or less wholeheartedly propel us towards those ends by whatever means it rationally takes to get us there.

Again, according to the Kantian theory of free agency that I'm expounding and defending here, the wholehearted self-realization of autonomous willing, or autonomous self-fulfillment, is *principled authenticity*. Every time an agent truly acts for the sake of the moral law, they realize absolute, intrinsic, non-denumerably infinite, obective moral value or moral worth, and thereby subjectively experiences an aspect of, or some salient degrees of, an ideally complete life of principled authenticity. They thereby experience deep happiness, self-fulfillment, or *Selbstzufriendenheit*. But if they *also* thereby subjectively experience achieves some individual or socially-shared shallow human happiness, then they *also* realize a proper part of the sole and complete good. Thus according to Kant's theory of the highest good, as I'm understanding it, there are *two* fundamental values or highest goods: (1) *the highest or supreme good*, namely, *the good will*, that is, principled authenticity, necessaruly including deep happiness, self-fulfillment, or *Selbstzufriedenheit*, and (2) *the sole and complete good*, represented by the moral regulative Idea of God, centered on human dignity, that is, individual and social *shallow* human happiness that's actively guided and controlled by a good will, and enacted by means of sufficiently respecting everyone's human dignity, which thereby expresses an ideal proportionality between moral virtue on the one hand and morally worthy happiness-as-self-fulfillment, together with shallow happiness, spread out over all of humanity.

In turn, the relation between the highest or supreme good and the sole and complete good is essentialist and mereological. An occurrently autonomous human real person's good will, or principled authenticity, enacted by means of sufficiently respecting everyone's human dignity, is the activating immanent structure (or "essential form") of the vital stuffing (or "prime matter") that is deep rational human minded animal happiness, and the living whole that is jointly constituted by them, propagated over all of rational humanity, is the sole and complete good.

Conclusion

In this essay, I've argued that our essentially embodied innate mental capacity for affective intelligence (i.e., desiring, emoting, and feeling) necessarily transcends the mechanical powers of any and all actual or really possible digital computing systems or digital technology more generally. As I noted in section 1, as I'm using this term and notion, affective intelligence is the same as our essentially embodied innate mental capacity for *caring*; and the particular mode of our capacity for caring that I'm focusing on is our essentially embodied innate mental capacity for *principled authenticity*. Explicitly spelled out in a nutshell, then, my argument that our essentially embodied innate mental capacity for principled authenticity necessarily transcends

the mechanical powers of any and all actual or really possible digital computing systems or digital technology more generally, is as follows.

1. As we've just seen in sections 2-3, the achievement of principled authenticity by rational human minded animals is a *creative transformation* that brings about a categorical improvement or upgrade in the intrinsic specific character or quality of the life of anyone who performs it.
2. All rational human minded animals are innately capable of achieving principled authenticity, at least to some salient degree and extent.
3. But according to what I've called *Babbage's Principle*, no digital computing system or digital technology more generally is capable of performing categorical improvements or upgrades in the intrinsic specific character or quality of the inputs, premises, or other materials provided to it (Hanna, 2025: section 2.7).
4. Therefore, our essentially embodied innate mental capacity for principled authenticity necessarily transcends the mechanical powers of any and all actual or really possible digital computing systems or digital technology more generally, *even when* we also fully allow for our finitude, fallibility, and thoroughgoing normative imperfection, as rational minded "human, all-too-human" animals.

Or to put the same conclusion more bluntly: henceforth, if any AI researcher, digital technology guru or true believer, or digital technology business corporation CEO or spokesperson, ever tells you that there is or ever could be really such a thing as "affective AI" "emotion AI," (see, e.g., Somers, 2019; Labbé, 2023), or "machines with emotional intelligence" (Picard, 2008), then you'll know that it's pure *bullshit* (Frankfurt, 1988d). Or, slightly less bluntly put, as Will Smith's character in the 2004 movie *I, Robot*, Detective Del Spooner, mockingly says to the rogue NS-5 robot Sonny when it fails to interpret the difference between a blink and a wink:[1] "It's a human thing. You wouldn't understand."

But, if we also mockingly say this to the bullshitting AI researchers, digital technology gurus or true believers, or digital technology business corporation CEOs or spokespersons, then it's a way of saying that they're all thinking, caring, and acting in a way that's fully *inauthentic*. Of course, in the context of the movie *I, Robot*, Spooner's mockery is lost on Sonny, since he/it is affectively vacuous. For better or worse, only rational human animals can be inauthentic: just as AI cannot be authentic, so too AI cannot be inauthentic.[2]

---

[1] This isn't to say that I'm philosophically endorsing the movie *I, Robot* as a whole. In fact, as the movie unfolds it's clear that—unfortunately—it ultimately buys into what I've called *the myth of artificial intelligence* (Hanna, 2024). But the opening scenes between the cynical, passionate Detective Spooner and the emotionally clueless and indeed affectively vacuous NS-5 robot Sonny are philosophically on-target.

[2] I'm grateful to the members of the Contemporary Kantian Philosophy Project, aka CKPP, for thought-provoking conversations on and around some of the main topics of this essay—especially the two-tiered, complex nature of happiness, and the nature of our innate capacity for achieving deep happiness, self-fulfillment, or

References

Budick, S. (2010). *Kant and Milton*. Harvard University Press.

The Contemporary Kantian Philosophy Project. (2023). *The 6th annual workshop of The Contemporary Kantian Philosophy Project: "Kant and the primacy of sensibility," Washington University in St Louis, USA, 19–20 May 2023*. https://www.academia.edu/41703363/The_6th_Annual_Workshop_of_The_Contemporary_Kantian_Philosophy_Project_Kant_and_the_Primacy_of_Sensibility_Washington_University_in_St_Louis_USA_19_20_MAY_2023_May_2023_version_including_names_of_participants_titles_and_abstracts_

Frankfurt, H. (1988a). *The importance of what we care about*. Cambridge University Press.

Frankfurt, H. (1988b). Freedom of the will and the concept of a person. In *The importance of what we care about* (pp. 11–25). Cambridge University Press.

Frankfurt, H. (1988c). The importance of what we care about. In *The importance of what we care about* (pp. 80–94). Cambridge University Press.

Frankfurt, H. (1988d). On bullshit. In *The importance of what we care about* (pp. 117–133). Cambridge University Press.

Frankfurt, H. (1988e). Identification and wholeheartedness. In *The importance of what we care about* (pp. 159–176). Cambridge University Press.

Frankfurt, H. (2004). *The reasons of love*. Princeton University Press.

Hanna, R. (2006). *Kant, science, and human nature* (The Rational Human Condition, Vol. 5). Clarendon/Oxford University Press. https://www.academia.edu/21558510/Kant_Science_and_Human_Nature

Hanna, R. (2015). *Cognition, content, and the a priori: A study in the philosophy of mind and knowledge* (The Rational Human Condition, Vol. 5). Oxford University Press. https://www.academia.edu/35801833/The_Rational_Human_Condition_5_Cognition_Content_and_the_A_Priori_A_Study_in_the_Philosophy_of_Mind_and_Knowledge_OUP_2015_

Hanna, R. (2018a). *Deep freedom and real persons: A study in metaphysics* (The Rational Human Condition, Vol. 2). Nova Science. https://www.academia.edu/35801857/The_Rational_Human_Condition_2_Deep_Freedom_and_Real_Persons_A_Study_in_Metaphysics_Nova_Science_2018_

Hanna, R. (2018b). *Kantian ethics and human existence: A study in moral philosophy* (The Rational Human Condition, Vol. 3). Nova Science. https://www.academia.edu/36359647/The_Rational_Human_Condition_3_Kantian_Ethics_and_Human_Existence_A_Study_in_Moral_Philosophy_Nova_Science_2018_

Hanna, R. (2022). *G.E. Moore's open question argument against the naturalistic fallacy revised, updated, and redefended* [Unpublished manuscript]. https://www.academia.edu/74336725/G_E_Moores_Open_Question_Argument_Against_The_Naturalistic_Fallacy_Revised_Updated_and_Redefended_March_2022_version_

Selbszufriedenheit via principled authenticity—at their 6th annual workshop in St Louis, in May 2023 (CKPP, 2023).

Hanna, R. (2023a). In defence of dignity. *Borderless Philosophy, 6*, 77–98. https://www.cckp.space/single-post/bp6-2023-robert-hanna-in-defence-of-dignity-77-98

Hanna, R. (2023b). *Dignity, not identity* [Unpublished manuscript]. https://www.academia.edu/96684801/Dignity_Not_Identity_February_2023_version_

Hanna, R. (2023c). *Frederick Douglass, Kant, and human dignity: A three-step argument that Kant's moral philosophy isn't racist* [Unpublished manuscript]. https://www.academia.edu/97518662/Frederick_Douglass_Kant_and_Human_Dignity_A_Three_Step_Argument_That_Kant_s_Moral_Philosophy_Isn_t_Racist_May_2023_version_

Hanna, R. (2024). The myth of AI, existential threat, why the myth persists, and what is to be done about it. *Borderless Philosophy, 7*, 35–61. https://www.cckp.space/single-post/bp-7-2024-robert-hanna-the-myth-of-ai-existential-threat-why-the-myth-persists-and-what-is-to

Hanna, R. (2025). *Digital technology for humans: The myth of AI, human dignity, and neo-Luddism*. De Gruyter Brill.

Hanna, R., & Maiese, M. (2009). *Embodied minds in action*. Oxford University Press.

Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). Blackwell.

Hume, D. (1978). *A treatise of human nature* (2nd ed.). Clarendon/Oxford University Press.

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Trans. & Eds.). Cambridge University Press.

Kierkegaard, S. (2000). *The essential Kierkegaard* (H. Hong & E. Hong, Trans.). Princeton University Press.

Labbé, A. (2022, November 23). Emotion AI: Why it's the future of digital health. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2022/11/23/emotion-ai-why-its-the-future-of-digital-health/

MacFarquhar, L. (2016). *Strangers drowning: Impossible idealism, drastic choices, and the urge to help*. Penguin.

Moore, G. E. (1903). *Principia ethica*. Cambridge University Press.

Oppy, G., & Dowe, D. (2021). The Turing test. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter ed.). https://plato.stanford.edu/archives/win2021/entries/turing-test/

Parfit, D. (2011). *On what matters* (Vol. 1). Oxford University Press.

Pascal, B. (1955). *Pensées* (A. J. Krailsheimer, Trans.). Penguin.

Picard, R. W. (2008). Toward machines with emotional intelligence. In G. Matthews, M. Zeidner, & R. D. Roberts (Eds.), *The science of emotional intelligence: Knowns and unknowns* (pp. 396–416). Oxford University Press. https://academic.oup.com/book/6088/chapter-abstract/149548563

Rilke, R. M. (1989). Archaic torso of Apollo (S. Mitchell, Trans.). In *Selected poetry of Rainer Maria Rilke* (pp. 60–61). Vintage Books.

Sarasohn, J. (1993). *Science on trial: The whistleblower, the accused, and the Nobel laureate*. St. Martin's.

Schapiro, T. (2009). The nature of inclination. *Ethics, 119*, 229–256.

Solnit, R. (2009). *A paradise built in hell: The extraordinary communities that arise in disaster*. Penguin.

Somers, M. (2019, March 8). Emotion AI, explained. *MIT Management, Sloane School*.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.

Wikipedia. (2026). *LaMDA*. https://en.wikipedia.org/wiki/LaMDA