JOURNAL OF APPLIED ETHICS AND PHILOSOPHY

Center for Applied Ethics and Philosophy
Hokkaido University

vol.13

February 2022

ISSN 1883-0129

Journal of Applied Ethics and Philosophy

Printed in Japan

ISSN 1883 0129 (Print)
ISSN 1884 0590 (Online)

All queries should be directed to:
The Editor-in-Chief
Center for Applied Ethics and Philosophy
Faculty of Humanities and Human Sciences
Hokkaido University
N10 W7, Kita-ku
Sapporo 060-0810
Japan
jaep@let.hokudai.ac.jp

CONTENTS

Editorial Note

The *Journal of Applied Ethics and* Philosophy is an interdisciplinary periodical covering diverse areas of applied ethics and philosophy broadly understood. It is the official journal of the Center for Applied Ethics and Philosophy (CAEP), Hokkaido University. The aim of the *Journal of Applied Ethics and Philosophy* is to contribute to a better understanding of ethical and philosophical issues by promoting research into various areas of applied ethics and philosophy, and by providing researchers, scholars and students with a forum for dialogue and discussion on ethical and philosophical issues raised in contemporary society. The journal welcomes original and unpublished regular academic papers as well as discussion papers on issues in applied ethics and philosophy broadly understood.

Nobuo Kurata
Editor-in-Chief

Many thanks to *Journal of Applied Ethics and Philosophy* reviewers

The *Journal of Applied Ethics and Philosophy* would like to thank the following individuals who most generously refereed manuscripts for us between February 2021 and January 2022. The assistance and expertise of these professionals promotes and maintains the high quality of the journals content. Thank you.

Kerrin Artemis Jacobs
Makoto Suzuki
Tomohiko Kondo
Shigeru Taguchi
Kengo Miyazono

The Evolution of Sartre's Concept of Authenticity

From a Non-Egological Theory of Consciousness to the Unrealized Practical Ethics of the Gift-giving (No-)Self

Lehel Balogh
Visiting Research Fellow, Center for Applied Ethics and Philosophy, Hokkaido University

Abstract

Over forty years have passed since the death of Jean-Paul Sartre, still, his oeuvre stands out as a paramount achievement in existential-phenomenological thought. Among the numerous ideas and challenges he offered to contemporary continental philosophy, the problem of authenticity deserves a special place, for it connects many of existentialism's key concerns. The ever reforming conceptualization of authenticity had spread from the mid-1930s (La transcendance de lʼégo) till Sartre's posthumously published Cahiers pour une morale that appeared in the early 1980s, and it had a profound impact not only on ethical and literary theories, but also on psychiatry and psychotherapy. The present essay's undertaking is to closely follow the trajectory of this celebrated concept, to contextualize its development in accordance with Sartre's shifting philosophical as well as ethical projects over the years, and to point out some of the affinities this concept might have with East Asian thought, and with Buddhism in particular

Keywords: authenticity, bad faith, existentialism, nothingness, self-transformation

> *Man makes himself; he does not come into the world fully made, he makes himself by choosing his own morality, and his circumstances are such that he has no option other than to choose a morality* (Sartre 1946/2007, 46).

When discussing the concept of authenticity in Sartre's thought, it is both customary and expedient to compare it to that of Heidegger's, the latter having had a palpable effect on Sartre's own meditations. Like Heidegger, Sartre also begins with the program of Husserlian phenomenology. Before probing into the deeply problematic questions of being, and particularly those of *Dasein*—which were indisputably central to Heidegger's early inquiries—, however, he turns to another key concept that was conspicuously absent from the German philosopher's reflections: *consciousness*. Already in his early essay, *The Transcendence of the Ego*, Sartre sets out to problematize those philosophically fundamental questions that will be elaborated in his later writings in more detail, and which will eventually lead to the radical conceptualization of authenticity. This early essay attempts to delineate the relationships between ego, self, and consciousness, and in its conclusion the orientations and boundaries of the mature Sartrean phenomenological ontology are also distinctively outlined.

What are the fundamental questions discussed in *The Transcendence of the Ego*? Most of these questions are linked, above all, to the nature of consciousness. In contrast to Husserl's view, Sartre claims that in order to explain the nature of perception, phenomenology does not need to postulate a transcendental consciousness which possesses egological structures. "Rather, consciousness is defined by intentionality" (Sartre 1936/2004, 3). To adequately describe how consciousness functions, we need nothing else but to point to the mental act of seizing objects; in other words, one must provide an explanation concerning the phenomenon of intentionality. According to Sartre, the mind is not a virtual space or container within which the ego—or the subject—resides. The primary activity of consciousness is a ceaseless self-transcendence toward the object: consciousness keeps stepping out of itself, so as to become the consciousness *of* an object. In itself, however, consciousness is nothing. Consequently, the notion that such entities as mental pictures or ideas exist in the mind is decidedly false: consciousness does not *store* anything—instead, it exists insomuch as it is the consciousness *of* something.

At this point we can see emerge one of Sartre's key ideas: namely, that the concept of an 'unconscious consciousness' is a self-evident self-contradiction. Consciousness is at all times conscious (of something), argues Sartre: it is always aware that itself is nothing else but consciousness. Nevertheless, "it becomes conscious of itself *insofar as it is consciousness of a transcendent object*" (Sartre 1936/2004, 4). But when is consciousness the consciousness of a transcendent object? According to Sartre, it has always been, and remains so, interminably. Consciousness is the uninterrupted consciousness of something *other than* itself, something that is thus beyond itself. By the same token, it is also the consciousness *of itself*, simultaneously—it is self-consciousness. How is this possible? The answer lies in Sartre's view that there are two distinct kinds of consciousnesses: an *unreflective* and a *reflective* one. In the first case, when the consciousness is unreflective, it is primarily the consciousness of a transcendent object, but 'non-thetically' it is the consciousness of itself, too. In other words, in this kind of unreflective consciousness, the object of consciousness is something external, but it does not entirely cease to be aware of itself, either—it does know about itself *without reflecting* upon itself. As soon as the direct object of consciousness becomes consciousness itself, unreflective consciousness turns into reflective consciousness. At that moment, the object of consciousness is nothing but consciousness itself.

"My question is this: is there any room for an I in a consciousness of this kind? The reply is clear: of course not" (Sartre 1936/2004, 5). If the ego has no place within consciousness, then it appears that perception as such—the emergence of concrete phenomena—does not take place due to the activity of personal cognition. (My) consciousness is not my consciousness; it precedes any sort of self-nature, personhood, or subjectivity.[1] Similarly to everything else that can appear on the perceptual horizon of consciousness, the ego is also transcendent to it. In other words, the ego does not belong to consciousness: it belongs to the world. Sartre draws attention to the fact that it is "consciousness that renders the unity and personality of my I possible. [Not the other way around— L.B.] The transcendental I thus has no *raison d'être*" (Sartre 1936/2004, 4). According to Sartre, the transcendental field of consciousness is characterized by spontaneity and impersonality. Individuality and subjectivity can only emerge 'outside', in the world, as a result of the reflective activity of consciousness, in relation to other people, and not 'inside' consciousness itself.

"Perhaps, indeed, the essential function of the Ego is not so much theoretical as practical … perhaps its essential role is to mask from consciousness its own spontaneity" (Sartre 1936/2004, 27). Sartre goes as far as to accuse consciousness for the creation of the ego 'as a false representation of itself'. The reason of this self-deception is that in pure (self)reflection consciousness recognizes itself as a limitless spontaneity and constant creative activity, and this confrontation with its own nature causes considerable anxiety: "It is this absolute and irremediable anguish, this fear of oneself, that in my view is constitutive of pure consciousness" (Sartre 1936/2004, 28).[2] This approach, in fact, effectively allows for the embedding of the Husserlian concept of 'natural attitude' along with the concept of '*epoché*' in an entirely new context. Inasmuch as it takes consciousness a specific effort to create a 'natural attitude' during the attempt of escaping from the anxiety that it had come

[1] It is noteworthy fact that Japanese philosopher Nishida Kitarō's characterization of 'pure experience' displays intriguing resonances with Sartre's descriptions of an egoless consciousness that Sartre identifies with 'nothingness'. Moreover, Nishida's account of direct experience which is also devoid of a personal agency and which is also portrayed as having to do with the acitivity of 'nothingness'—although in a rather different sense—gives further reason for comparisons. As Nishida famously wrote, some twenty years before Sartre stated that it is consciousness that establishes the appearance of the individual self: "It is not that there is experience because there is an individual, but that there is an individual because there is experience" (Nishida 1911/1990, 19). For more on Nishida's and Japanese philosophy's connections and potential influences on the young Sartre see Light 1987.

[2] This observation is echoed by Japanese philosopher and psychiatrist Bin Kimura who contends that when looking for secure spots by which the self could anchor itself in reality and gain comfort and reassurance in the face of the unsettling fluidity of being, the self only manages to entangle itself deeper and deeper into a false representation of reality. This false representation of the world attempts to flee from the inherently event-like nature of phenomena by turning them into things that appear stable; nonetheless, the stability they seem to provide is a fake one. The self compulsively tries to escape from the awareness of the volatility of the world but the world repeatedly informs the self that the efforts on the part of the self to stabilize itself and the world by way of transmuting them into unchanging entities are, after all, spurious (Kimura 1982/2011).

face to face with while discovering its own intrinsically spontaneous nature, the attributive adjective 'natural' in 'natural attitude' does not seem so natural after all as it had appeared before, in Husserl. In addition, the phenomenological *epoché* does not seem to be such a purely and rigorously scientific procedure either, the way Husserl liked to depict it; rather, according to Sartre, it "is an anguish that imposes itself on us and that we cannot avoid" (Sartre 1936/2004, 28).

One's self, then, regarding which one would intuitively assume that it *absolutely* belongs to one, that one can trust it without hesitation and that it can define one precisely and correctly, well this self, according to Sartre, is neither absolute nor reliably stable. Instead, it is, in fact, nothing more than a mere 'solution' to an existential problem, created in 'bad faith' (*mauvaise foi*) by our consciousness so that it could hide from itself and flee from the experience of existential anguish and anxiety. The self that belongs to me is not, in any way, more dependable than the selves that belong to others. "My I, indeed, *is no more certain for consciousness than the I of other men*. It is simply more intimate" (Sartre 1936/2004, 29). The self, as mentioned earlier, is 'outside' in the world among other beings. Nonetheless, man for Sartre is essentially identical with consciousness—or, as consciousness is called in *Being and Nothingness*: *être-pour-soi*, 'being-for-itself'. As a consequence, the main undertaking of the Sartrean *magnum opus* is precisely the analysis and accurate description of this entity: *man as a being-for itself*.

According to the Sartre of *Being and Nothingness*, entities can be classified into two major categories: 'being-in-itself' (*être-en-soi*) and 'being-for-itself' (*être-pour-soi*). These two are associated with and inhabit two ontologically separate realms. Having said that, they are nevertheless interdependent: neither exists without the other, neither can be derived from the other. The 'being-in-itself' in Sartre is the concrete phenomenon which stands opposed to the 'being-for-itself', that is, consciousness. Together they make up the synthetic unity of consciousness and phenomenon. 'Being-in-itself' is characterized by the fullness of being and is complete positivity: "being is what it is". It contains no negation, has no distance from itself, and, accordingly, it has no relation to itself, either: therefore, it is solidly self-identical. By contrast, 'being-for-itself' is pure spontaneity, which "can always pass beyond the existent, not toward its being, but toward the meaning of this being" (Sartre 1943/1978, lxiii). Sartre already foreshadows in the introduction of his book that the 'being-for-itself' cannot possibly coincide with itself, for it exists only in a continuous movement that keeps drawing it *away* from itself. If 'being-in-itself' is "being what it is", then, conversely, 'being-for-itself' is "being what it is not" (Sartre 1943/1978, lxv).

As a result, insofar as 'being-for-itself' is everything that the self-identical, solid being is *not*—keeping in mind that consciousness actually does not even possess a self-identity—, there remains no other alternative open for Sartre than to postulate consciousness as that which stands in complete opposition to being: that is, consciousness is non-being or *nothingness*. To rephrase it, nothingness is essentially identical with consciousness or 'being-for-itself'. More precisely, nothingness is the activity of human existence through which a fissure occurs in the texture of being. Nothingness cannot originate from the 'being-in-itself', for 'being-in-itself' connotes a complete fullness of being. In other words, "if being is everywhere, it is not only Nothingness which, as Bergson maintains, is inconceivable; for negation will never be derived from being. The necessary condition for our saying *not* is that non-being be a perpetual presence in us and outside of us, that nothingness haunt being" (Sartre 1943/1978, 11).

As Kalmanson points out, "Sartre deals with alienation throughout *Being and Nothingness* in his exploration of the question of how the subject gains access to anything outside of its own awareness. Ultimately, his solution hinges on the idea of non-being" (Kalmanson 2021, 112-113). Negation, the possibility—and actuality—of motion and change enters into the world in virtue of the particular activity of the consciousness which Sartre calls "nihilation" (*néantisation*). The existence of consciousness entails practically nothing but this activity. With this explication Sartre indicates that consciousness is the only (non) being which is able to breach and penetrate the otherwise wholly homogenous, rock-solid walls of being by virtue of creating distinctions, of negating and denying being, hence essentially by being able to break away from the givenness of the actual reality of any concrete situation.

Having arrived at this point, it is only one more step in the logical chain to claim that consciousness is responsible not only for bringing into the world the ability of transcending reality, but also for creating the possibility for freedom. According to Sartre, consciousness in itself represents boundless freedom, because it is in the individual's power to supersede and supplant any concrete phenomenon at one's will: as soon as man ceases to relate to a given set of phenomena, those phenomena will, in turn, also cease to relate to man. One could say that this is, then, an unequal relationship, for establishing and maintaining this relation between consciousness and phenomena depend exclusively on man. This is why Sartre regards the freedom of the 'being-for-itself' boundless because one can decide whether one wishes to engage the phenomena or, instead, negate them. As the oft-quoted thesis of Sartrean existentialism declares:

> Human freedom precedes essence in man and makes it possible; the essence of the human being is suspended in his freedom. What we call freedom is impossible to distinguish from the *being* of "human reality." Man does not exist first in order to be free *subsequently;* there is no difference between the being of man and his being-free (Sartre 1943/1978, 25).

As one can observe, Sartre equates human reality with human existence, as well as with consciousness and with 'being-for-itself'. All these titles and labels designate the uniquely human ontological status whose nature is that it has *no nature* whatsoever. Freedom is not just one of the accidental attributes of man, but it *is* human existence itself. Man—neither as a genus, nor as an individual—has no prewired, rigid frame of ontological structure that would predestine its existence. The individuals' destinies are to autonomously create themselves from scratch, without the reassuring authority of a tradition on which one could lay all the blame if things happen to go astray.[3] The extent to which one attempts to escape from one's personal freedom and take refuge in prefabricated constructs—such as the belief in an unalterable personal destiny or in a fixed individual personality or in absolute goals and objective values—indicates the level of denial and self-deception vis-á-vis the emptiness and fundamental meaninglessness of one's existence. "Freedom is the human being putting his past out of play by secreting his own nothingness … consciousness continually experiences itself as the nihilation of its past being" (Sartre 1943/1978, 28). Consciousness perpetually breaks away: not just from the world, but from itself, from its own past too; and its freedom is constituted in this very act of incessant disengagement and breaking away.

That said, as we already learned from the reasoning of *The Transcendence of the Ego*, consciousness tends to flee from the consciousness of its *own* freedom, that is, from itself as 'freedom-consciousness'. The reason for this, as Sartre reiterates in *Being and Nothingness*, is that the awareness of its infinite freedom fills one with anxiety: "it is in anguish that man gets the consciousness of his freedom, or if you prefer, anguish is the mode of being of freedom as consciousness of being; it is in anguish that freedom is, in its being, in question for itself" (Sartre 1943/1978, 29). 'Anguish' in Sartre's thought plays a similarly crucial role as *Angst* in Heidegger's existential analytic: just as in *Being and Time*, nothingness manifests itself through the experience of anguish/anxiety. In this scenario one has no choice but to face up to one's own nothingness, to one's groundless existential ground. But, whereas 'nothingness' for Sartre is actually indistinguishable from the human individual and 'nihilation' is the activity of human consciousness, the same cannot be claimed about Heidegger's concept of nothingness. Both authors, however, would agree with Kierkegaard on the elementary difference between fear and anguish/anxiety in the sense that while the latter has no definite object, the former is always directed toward a concretely circumscribable object; when we fear, we always fear *something*.

Sartre maintains that while experiencing anguish the individual is anxious about their personal existence as well as about their total freedom and complete responsibility, too. One is anxious either about one's past or about one's future. If we are anxious about our past, "anguish appears as an apprehension of self inasmuch as it exists in the perpetual mode of detachment from what is" (Sartre 1943/1978, 35). Nothing, not even one's own past can determine what one will or what one should do next. Naturally, man must always become something or somebody, since even if one is reluctant to make strong and concrete commitments, this noncommittal attitude itself is the result of an original—perhaps unreflected—decision: a decision not to be committed. Thus, ultimately, one must always decide with every single moment who one is and who one should become. On the other hand, when one is anxious about the future, one comes to realize that "I am not the self which I will be" (Sartre 1943/1978, 31). Just as the past is unable to determine what the present would become, similarly, the present is unable to force the outcome of the future, argues Sartre. Since man exists in a constant displacement and becoming, one can never completely coincide with oneself. Yet, according to Sartre, man still steadfastly holds onto this belief: by positing a self-identical ego and attributing it to oneself, one proceeds, in effect, to *deliberately* deceive oneself. This self-deceit cannot be understood to constitute an innocent or naïve sort of misconception: it is a perfectly conscious move, since consciousness, as we may remember Sartre's earlier argument, can only perform conscious actions—consciousness is unable to *unconsciously* deceive itself and to hide this deception from itself. When consciousness, in fact, deceives itself, this kind of self-deception is for what Sartre applies the term 'bad faith'.

Concerning the strategies people employ in bad faith, we often try to flee from anguish „by attempting to apprehend ourselves from without as an Other or as a *thing*" (Sartre 1943/1978, 43). We thus quite literally objectify ourselves. To put it differently, we artificially freeze ourselves into a pretense that makes us believe that we have, indeed, become such ontologically solid

[3] This is a point where Sartre and Heidegger disagree strongly: the significance of a tradition for authentic existence. For Heidegger, no authentic Dasein can exist without the reappropriation of its tradition; for Sartre, no authentic existence can exist with it.

objects that have stable, fix, 'objective' natures. Of course, we know all too well, that we cannot actually possess such an object-like ontological makeup, adds Sartre. We flee from the inescapable freedom bestowed upon us, yet in this flight we are never truly capable of forgetting about that which we have been trying to get away from in the first place: "the flight from anguish is only a mode of becoming conscious of anguish. Thus anguish, properly speaking, can be neither hidden nor avoided" (Sartre 1943/1978, 43).[4] Man, by nature, is a free being, maintains Sartre; nevertheless, this position also implies that man, by nature, is a being forever in anguish. In other words, these two, freedom and anguish, are inseparably linked together. 'Bad faith' enters into the scene in that particular dialectical relationship which one would try to escape by denying one's existential anguish, along with one's freedom and the ability for nihilation. One is unable to successfully carry out this denial, however, for in order to be able to escape and deny, one must already possess the ability to separate oneself from one's anguish; that is, one must already possess the ability of nihilation. To put it differently, 'bad faith' is aware of its own bad faith and its own self-deception at all times: that it runs away from its freedom, and that it fails to take responsibility even for this running away.

Nevertheless, taking responsibility becomes unavoidable once one owns up to one's prior deliberate self-deceptions, and, instead, confronts the contingency of one's existence. Contrary to widespread popular belief, Sartre did not argue that since there are no objective values in the world, then everything is morally permissible. What he did, in fact, assert regarding the nature of values was the following: "Anguish is opposed to the mind of the serious man who apprehends values in terms of the world and who resides in the reassuring, materialistic substantiation of values" (Sartre 1943/1978, 39). Values cannot simply reside somewhere outside in the world *a priori* given to us, Sartre asserted, because in that case they would cease to be *valuable*; as objective entities they would instantly lose their worth. A value is valuable only inasmuch as it has relevance in relation to man's existence: that is, only if the meaning and significance which the value represents is not an *absolute* meaning and significance, but a particular one which then must always be situational, too. A true value can only be constituted in the recognition created by that 'active freedom' which, according to Sartre, is identical with the human agent:

> It follows that my freedom is the unique foundation of values and that nothing, absolutely nothing, justifies me in adopting this or that particular value, this or that particular scale of values. As a being by whom values exist, I am unjustifiable. My freedom is anguished at being the foundation of values while itself without foundation. It is anguished in addition because values, due to the fact that they are essentially revealed to a freedom, can not disclose themselves without being at the same time "put into question," for the possibility of overturning the scale of values appears complementarily as my possibility (Sartre 1943/1978, 38).

Every moment man is 'thrown into the world', which means that the postulation of values does not take place only on a theoretical level but values are being created continuously through real-life choices and practical decisions. Since one must create and realize the meaning of one's existence—along with the meaning of the world—, the values are actually upheld by the individual; consequently, one truly is what one makes of oneself. "In anguish I apprehend myself at once as totally free and as not being able to derive the meaning of the world except as coming from myself" (Sartre 1943/1978, 40).

Is 'bad faith' or inauthenticity (the two are practically the same term in Sartre) avoidable? As Sartre points out, "the first act of bad faith is to flee what it can not flee, to flee what it is" (Sartre 1943/1978, 70). Inauthenticity therefore lies primarily in the fact that one denies who one really is. Man in Sartre's opinion—just as in Nietzsche's—is essentially a value creator, a freedom of existence that lives in a contingent world bereft of inherent structures of meaning. Perhaps one would find it reasonable to suggest that the opposite of 'bad faith', i.e. 'good faith' and honesty could pave the way toward authenticity. Sartre's response to this suggestion is clearly in the negative. "The ideal of good faith (to believe what one believes) is, like that of sincerity (to be what one is), an ideal of being-in-itself" (Sartre 1943/1978, 69); therefore, it does not concern the existence of the being-for-itself. Consciousness can never become a 'being-in-itself'; it would be in vain trying to flee from the anguish of having to face freedom, escaping toward transcendence, toward a comforting tranquility of being. A human individual is unable to consolidate itself into an unchanging entity in 'good faith'. In this way, Sartre demonstrates that the idea of 'good faith' is no better and no more tenable than the idea of 'bad faith': both refer to inauthentic forms of the human existence.

> Good faith seeks to flee the inner disintegration of my being in the direction of the in-itself which it should be and is not. Bad faith seeks to flee the in-itself by means of the inner disintegration of my being. But it denies this very disintegration as it

[4] This might remind one of Kierkegaard's contention that the more one tries to escape from existential anxiety, the more one will get oneself entangled in it.

> denies that it is itself bad faith (Sartre 1943/1978, 70).

Even so, it is not entirely inconceivable to step out of and leave behind the *circulus vitiosus* of inauthentic modes of existence in Sartre; it is, in fact, possible to „radically escape bad faith. But this supposes a self-recovery of being which was previously corrupted. This self-recovery we shall call authenticity" (Sartre 1943/1978, 70, footnote). The detailed elaboration of authenticity was not among Sartre's projects during the writing of *Being and Nothingness*; he reserved this undertaking for his later works. *Being and Nothingness*, he emphasized, could only deal with the task of what *is*, and not with what *should be*, as the book was, strictly speaking, ontological in nature: that is, it was supposed to be about what exists. "It does, however, allow us to catch a glimpse of what sort of ethics will assume its responsibilities when confronted with a human reality in situation" (Sartre 1943/1978, 625-626). Regarding the nature of human existence, Sartre has come to expose and portray it as an essentially inexhaustible desire for total satisfaction. This total satisfaction can never become an actual reality: the 'being-for-itself' can never transform itself into becoming 'being-in-itself'. The innermost desire of consciousness, according to Sartre, is to have the same sort of solid self-identical nature that the 'being-in-itself' possesses. At the same time, however, consciousness wants to retain its own absolute freedom as well. This seeming contradiction can neither be resolved, nor removed from the scheme. Incidentally, this is why values actually emerge, since a value is a sort of ideality which reinforces the fantasy of the aforementioned unattainable synthesis: of becoming an entity with solid self-identity *and* to keep one's absolute freedom, too. According to Sartre, an 'existential psychoanalysis' would need to undertake the project of examining the universal human pursuit that attempts to achieve this synthesis whereby the inauthentic individual might become authentic.

'Existential psychoanalysis' in Sartre's nomenclature is not a purely scientific or medical practice, but instead a "*moral description*, for it releases to us the ethical meaning of various human projects" (Sartre 1943/1978, 626.). 'Human projects' are those personal plans via which one projects oneself in the future in the form of an attainable chosen goal—or what amounts to the same, to work toward a chosen value. "But the principal result of existential psychoanalysis must be to make us repudiate the *spirit* of seriousness" (ibid.). Why would such an attitude be advantageous? The 'spirit of seriousness', according to Sartre, possesses two characteristics that promote the operation of 'bad faith'; "it considers values as transcendent givens independent of human subjectivity, and it transfers the quality of 'desirable' from the ontological structure of things to their simple material constitution" (Sartre 1943/1978, 626). In sharp contrast to this, existential psychoanalysis ventures to transcend the 'spirit of seriousness' by virtue of revealing the moral agent of the human actions. The moral agent is "the *being by whom values exist*. It is then that his freedom will become conscious of itself and will reveal itself in anguish as the unique source of value and the nothingness by which the world exists" (Sartre 1943/1978, 627).

Considering all these claims together, one may well wonder how one could possibly become an authentic self in Sartre. What would await an individual if they were to renounce all forms of 'bad faith' and would choose to face up to existential anguish instead?

> A freedom which wills itself freedom is in fact a being-which-is-not-what-it-is and which-is-what-it-is-not, and which chooses as the ideal of being, being-what-it-is-not and not-being-what-it-is. This freedom chooses then not to recover itself but to flee itself, not to coincide with itself but to be always at a distance *from* itself (ibid.).

It would be hard to exaggerate the importance of this crucial passage in terms of its significance in the evolution of Sartre's concept of authenticity. This is the point where Sartre makes it unequivocally clear that the idea of authenticity envisioned by him is radically different in comparison to what past existentialist thinkers, namely Kierkegaard and Heidegger, have described. According to Sartre, bad faith or inauthenticity is maintained by the spirit of seriousness which can only be eliminated if one does not even *try* to actively transform one's inauthentic self to become a truly genuine self (or to 'win oneself', as Kierkegaard would have us do). These and similar efforts would only achieve to thrust one back and to plunge the individual amidst the tireless waves of the ocean of bad faith. Humans are authentic only insofar as they actively will themselves to be free, that is, if they decide *not* to want to coincide with themselves. In some sense, this suggestion appears somehow to be akin to Nietzsche's idea of authenticity in that it also strongly criticizes the generally accepted axioms concerning the existence and the supposed enduring nature of a personal character. Then again, Sartre perhaps goes even further than Nietzsche when he emphatically calls for avoiding even the *appearance* that it would be somehow possible to create a well-defined nature—a 'style' or a 'taste', as Nietzsche would put it— for ourselves in an authentic fashion. We are irredeemably contingent beings for Sartre that are always in the process of becoming something and somebody else; hence, the acknowledgement of this predicament, along with assuming the responsibility for the consequences is the only workable way for Sartre that could lead to authenticity.

By following Sartre's reasoning, one wonders whether all the above amounts to that the personhood of an individual would inevitably disintegrate to escape bad faith and become authentic. If there is not an unshakable integrative core of the personality to speak of, then what or who could be held accountable for the actions of the individual? If there is no self-identity, who can be identified as the actor of the actions? It is important to stress that Sartre never stops insisting that every person is completely responsible for the consequences of their actions as well as of what they make of themselves. What one makes of oneself is what one actively does, what one commits oneself to: an individual is nothing but the accumulated aggregate of their choices and actions. However, nothing obliges or determines one to choose one's particular obligations. There exists neither justification, nor valid external assistance that could tell one what to do: how to live one's life. Even the interpretation and explanation of the events that take place during one's lifetime is a task that only the self entitled to carry out; nobody else can take the burden away of having to give meaning and significance to the events of our lives.

In the last sentences of *Being and Nothingness*, Sartre anticipates the undertaking of a future ethical work that would build upon the findings of his lengthy ontological reflections. This ethical work, which would (have) surveyed the possibility of authenticity, was never published during his lifetime, although many hundreds of pages have been completed and compiled throughout the years. Nevertheless, he did not wish to make them public. Thus the *Cahiers pour une morale* could only appear posthumously, in 1983. In this philosophical diary Sartre worked on the detailed elaboration of the previously envisioned ethical work which he announced at the conclusion of *Being and Nothingness*. While studying the highly original musings of the *Notebooks for an Ethics*, one cannot help but marvel at the intensity and the tremendous intellectual effort and struggle with which Sartre endeavored to harmonize the concepts of morality, particularly that of responsibility, with the almost contourless character of human consciousness. According to the program of this ethical undertaking, inauthentic consciousness should reach the state of pure reflection[5] so as to be able to radically move away from bad faith and to recover the being that was corrupted by inauthentic consciousness itself.

> The passage to pure reflection must provoke a transformation:
>
> of my relation to my body. Acceptance of and claiming of contingency. Contingency conceived of as a chance.
>
> of my relation to the world. Clarification of being in itself. Our task: to make being exist. True sense of the In-itself-for-itself.
>
> of my relation to myself. Subjectivity conceived of as the absence of the *Ego*. Since the Ego is εξις (psyche).
>
> of my relation to other people (Sartre 1983/1992, 12).

A considerable shift can be detected here in Sartre's conception of human existence compared to that of *Being and Nothingness* which was rather inclined to reduce human reality to the absolute freedom of consciousness's nihilating potential, and to marginalize the individual's facticity—such as one's body or past—as non-essential parts of one's existential composition. The *Notebooks*, conversely, defines man as a being which can only be a being-for-itself due to the being-in-itself character of its bodily existence that connects it to the world. In other words, man's transcendence is based on man's facticity:

> In every perception of a thing I understand myself as a thing. I apprehend my own passivity along with the weight of this stone (I am what it weighs upon) but this passivity is at the same time a form of activity (I raise my hand, I move the stone from this place to that). A perpetual double relation. I could not act if I were not passive. Yet I can only be passive because I act (otherwise, I would just *be*, that is all); I am that being who through passivity and activity comes into the world for the In-itself and for myself. Passivity is *my connection* to the In-itself, both an ontological and a practical connection at the same time (Sartre 1983/1992, 51-52).

The being-in-itself can neither be passive, nor active: it simply exists. Contrarily, the sole reason why humans are able to be active is that they can be passive, too: it is passivity that allows them to comprehend the phenomena of the world. Bodily existence conceived as part of one's facticity has thus achieved a noticeably higher position in the *Notebooks* in comparison with its place in previous Sartrean works. The role of the body has been indeed reevaluated but it has not become overrated. Personality is seen by Sartre to be made up both by the 'situation'—

[5] Once again, the parallel between Nishida's concept of pure experience where the egoless vision of an authentic no-self can emerge and Sartre's conceptualization of pure reflection where bad faith and the self could finally disappear are remarkable.

which has been determined by one's facticity—and by the future projections of the consciousness; separately viewed, these are mere abstractions. Individuals cannot detach themselves from their specific historical situatedness, and in extreme cases even the freedom of choice could seemingly vanish from their lives. For instance, a terminally ill person cannot simply choose not to be terminally ill. Notwithstanding, one's attitude towards the concrete situation, according to Sartre, will always belong exclusively to one's personal area of competence: it is up to the individual whether they would choose to despair because of their looming death, or, on the contrary, whether they would gather the courage to boldly confront it. One is always free to pick one's attitude against a concrete situation.

For Sartre, there is no such thing as 'human nature' which would command us with absolute authority how to react in a given situation. This view can also be bolstered by acknowledging that concrete situations are not uniform occurrences of general 'basic situations', but are through and through singular cases. On this basis, one could claim that there is not a single experience or situation concerning dying that would be identical to any other experience or situation concerning dying; all of these are singular and unrepeatable events that belong to the lives of completely unique and singular individuals.

For that reason, the *Notebooks* argues that we do not need an abstract ethics but rather a concrete one which takes the present historical situation as its starting point and which places particular sets of goals in front of the individual existences, while discounting the pursuit after universal values. Human existence in its 'natural attitude' is forever working on creating a being-in-itself from itself. This effort, as we have learned it from the pages of *Being and Nothingness*, is not a viable project, however. What humans strive for basically is to construct a god out of themselves: a god that is able to do and achieve everything the self desires. As soon as one acknowledges the impossibility of this dream, though, one finally enables oneself to realize the experience of pure reflection upon which Sartre theorized already in *The Transcendence of the Ego*.

Pure reflection indicates the beginning of a new way of looking at things, a new vision, a novel approach. The term 'reflection' however can be a bit misleading here: it is not merely a disinterested contemplation, but rather a new venture and a vision that is motivated by the attainment of a particular goal. Pure reflection represents a positive, practical, and realistic morality which aims to realize an ever growing degree of freedom. Whereas, according to Sartre, "Being and Nothingness is an ontology before conversion" (Sartre 1983/1992, 6), the *Notebooks* endeavors to provide an account of the experience that is characteristic of pure reflection *after* the conversion.

The experience of the post-transitive attitude is the experience of *authenticity*. In pure reflection, one accepts the fact that one is not a necessary, substantive entity that has an unquestionable right to exist, but instead an unjustifiable and contingent freedom of existence that endlessly asks itself about the meaning and goals of its existence. One comes to accept that there are not any pre-given, *a priori* values or an authority to which one could appeal for the validation of one's life. Insofar as the individual conceives itself to be a free, contingent, and unjustifiable being, it transcends the dialectics of 'good faith–bad faith': in this way, one can at last realize a truly authentic existence. Yet, Sartre warns us: "If you seek authenticity for authenticity's sake, you are no longer authentic" (Sartre 1983/1992, 4).[6] Being authentic is not a value or a goal in itself but the incidental fruit that accompanies our having been able to come to terms with the groundlessness of our existence and also with the inescapable existential condition that we must create ourselves out of nothingness. This creation, however, is not the construction of something permanent and secure; on the contrary: it is the building of an existence, a (no-)self that is forever in flux, for which one takes full responsibility all the same.

Sartre argues that the world does not have a prearranged, original, objective meaning structure. This, though, does not mean that we would not be able to give legitimate meanings to things. Man is that particular being whose task precisely is to bring meaning into an otherwise perfectly meaningless universe, and this meaning must be constructed in a way that it has significance to one's life practice. The world, as long as we understand it as the already existing multiplicity of discreet objects, has not been created by the activity of human consciousness, evidently. By contrast, if we understand the world as an *orderly* diversity, then this orderly fashion could have only been brought about among the wealth of phenomena by the organizing activity of human consciousness.[7] Furthermore, the human individual, being the nihilating consciousness that it essentially is, has already always been the creator of the meaning and sense of the world (a recognizably Heideggerean thought), and even if one decided to opt out from being a creator of the meaning-universe, one would never actually succeed in doing so. According to Sartre, we are 'condemned to freedom', and by the same token, we are condemned to eternal creation as well. An

[6] This caveat reminds us of Zen Buddhism's guidance: if one sits down to meditate with the deliberate aim of becoming enlightened, one has already missed the chance of realizing enlightenment. Only by not trying to achieve, can one achieve it.

[7] This is precisely what Camus stated in *The Rebel*: man must create order out of the chaos and absurdity that is the *world*.

authentic personality must then affirm not only one's freedom but also one's own arbitrary creative activity. In other words, only those individuals can become authentic that learn to establish meanings amid the wholly chaotic indifference of being.

Attributing meaning to the world and to our lives is only one side of the coin. Carved on the other side is the resignation of the individual that makes one to decide to get rid of all of one's previous efforts that aimed to form the self into a being-in-itself-style substance. That is to say, one must give up voluntarily and 'generously' the project of becoming oneself so that this 'giving up' may become a 'giving to', a *gift*: a giving of meanings and values to the things that surround and envelope us. Just as in Nietzsche where man is identified as the creator of values, in Sartre too, the extraordinary task—which is at the same time the heaviest burden of human existence—is the necessity of having to give values and assign meanings to things. In light of this, for Sartre, moral behavior becomes in essence the act of gifting ourselves away, of elevating self-presenting into a general practice: it is an "absolute generosity, without limits, as a passion properly speaking and as the only means of being. There is no other reason for being than this giving. And it [is] not just my work that is a gift. Character is a gift" (Sartre 1983/1992, 129.). It is true even more so considering that being can only be 'saved' and transformed into something more humane—that is, more valuable, more ethical, and, ultimately, more free—if authentic individuals establish their own creative freedom as the foundation of the world. Sartre believes that the consequence of this would be that with such an attitude people would experience, for the first time in their lives, what it feels like to be the basis of an existence and a world that was shaped solely by their own efforts and creative powers. This would, in turn, also mean that they would immediately cease to be unjustifiable and contingent entities that aimlessly ramble about in a meaningless, hectic universe. Authentic individuals—due to the circumstance that they attribute meaning and values to the world on their own and also because they affirm themselves as the creators of this world that is abundant in meanings—would become morally superior to inauthentic individuals who try relentlessly, to no avail, to flee from their own creative freedom.

For Sartre, man is not god; creating a world *ex nihilo* is not in mankind's power. In a certain sense, however, it is appropriate to call man's freedom to create 'absolute'—insofar as it is not dependent on anyone or anything else. A human is a creator, a creator of values, simply because human existence, as we have seen, endows man with the task of creation. It is possible to subjectively justify the values and the meanings of the world, but "I can never persuade Others of my objective necessity … It is me, which nothing justifies, who justifies myself inwardly" (Sartre 1983/1992, 482). It appears, then, that the idea of subjective necessity is satisfactory for Sartre to argue that the ideal of freedom ought to stand on the top of the moral hierarchy of values. The reason why freedom must be the *foremost* value is that the freedom of values can only rest on the basis of acknowledging the principal value of freedom itself. Were it not in my freedom to freely evaluate what is valuable and what is not, how could we even talk about freely chosen values in the first place? This provides the key to understanding why Sartre stressed in his famous presentation *Existentialism is a Humanism* that by choosing one's own freedom one chooses, at the same time, the freedom of others, as well. If one did not do so, one would slit the roots of one's very own freedom (Sartre 1946/2007).

After having arrived at this point, it appears uncomplicated now to integrate the elements of interpersonality—which was portrayed in *Being and Nothingness* as a rather problematic issue that would significantly contribute to the alienation of the self—into the fabric of an authentic existence. As a matter of fact, Sartre does not only use the term 'authentic' to describe the desirable manner in which an individual should inhabit its own existence but also to illustrate one's ideal form of relationship to others. In this sense, a human relationship can also be conceived as authentic, provided that one acknowledges that the totality of one's human existence includes not only one's mode of being-for-itself (*freedom*) but also one's mode of being-in-itself (*facticity*): that is, one's past, and, likewise, one's body. In pure reflection, i.e. following the existential conversion, the objective aspect of human existence, which is readily accessible for other people as well, does not bring about automatic alienation from the self or estrangement from the others (or from the world). Sartre expands:

> This comes about only if the Other refuses to see a freedom in me *too*. But if, on the contrary, he makes me exist as an existing freedom as well as a *Being/object*, if he makes this autonomous moment exist and thematizes this contingency that I perpetually surpass, he enriches the world and me, he *gives a meaning* to my existence *in addition* to the subjective meaning I myself give it, he brings to light the *pathetic* aspect of the human condition, pathos I cannot grasp myself, since I am perpetually the negation through my action of this pathos (Sartre 1983/1992, 500).

The generosity, then, with which authentic individuals decide to give up the project of morphing themselves into complete being-in-itself-style entities—which, as we know, is an impossibility in any case—and instead give meanings and values to the world, finds its full and final form in the gesture that secures the existence and the freedom of the other person as a being-for-itself as well.

We may conclude our overview by stating that in Sartre the meaning and significance of the authentic individuality extends beyond itself. Although their goals are essentially particular to their personalities and life projects, authentic individuals are also universal inasmuch as a chief objective of theirs is to assist their fellow human beings in the realization of *their* goals. Sartre mentions 'authentic love' and 'authentic friendship', too, by which he, in my interpreation, means something very similar to the later Heidegger's notion of 'letting beings be' (*Gelassenheit*). Authentic love for Sartre does not attempt to arbitrarily solidify the reality of the other person into objective categories; instead it lets the other person's freedom come forth and freely reveal itself. In other words, authentic modes of relating do not try to shape the unfolding personal reality of the other to the self's own image. Authentic individuals would not try to coerce the ever-changing, living, event-like reality of the other into inflexible mental images that would inevitably distort the unique and fluid reality of another person. One should keep in mind, though, that Sartre does not speak about the letting be of *any kind* of freedom: he refers exclusively to the freedom of the authentic personality that has already gone through the transition that was brought about by the renunciation the workings of bad faith. Only this kind of personality who has migrated to a higher ethical plain and acquired a clearer vision of reality has been enabled, for Sartre, to support the plans and projects of other authentic individuals, since, evidently, not every plan is worthy of the authentic person's support. Those projects, however, that indeed deserve encouragement and aid, must be appropriately acknowledged; supporting them is, in fact, a moral obligation for the authentic individual. The willingness to allow and assist the other persons to attain their own authentic freedom, for instance, constitutes such a moral obligation.[8] In Sartre, this sort of support could only become realistically possible if the other person is also willing to grant me my own existential freedom; that is to say, if freedom is recognized between equal parties in a mutually respectful manner.

[8] Similarly, in Mahāyāna Buddhism the willingness to assist others to find liberation after one has achieved liberation for onself is an ethical imperative which is freely realized and enacted by the enlightened person (cf. the Boddhisatva way).

References

Kalmanson, L. (2021), *Cross-Cultural Existentialism: On the Meaning of Life in Asian and Western Thought*, London/New York: Bloomsbury.

Kimura, B. (1982/2011), *Time and Self*, Translated by John W. Krummel, in James W. Heisig, Thomas P. Kasulis and John C. Maraldo (eds.), *Japanese Philosophy: A Sourcebook*, Honolulu: University of Hawai'i Press, 958-972.

Light, S. (1987), *Shūzō Kuki and Jean-Paul Sartre*: *Influence and Counter-Influence in the Early History of Existential Phenomenology*, Carbondale and Edwardsville: Southern Illinois University Press

Nishida, K. (1911/1990), *An Inquiry into the Good*, Translated by Masao Abe and Christopher Ives, New Haven and London: Yale University Press.

Sartre, J-P. (1936/2004), *The Transcendence of the Ego: A Sketch for a Phenomenological Description*, Translated by Andrew Brown, London and New York: Routledge.

Sartre, J-P. (1943/1978), *Being and Nothingness: A Phenomenological Essay on Ontology*, Translated by Hazel Barnes, New York: Pocket Books.

Sartre, J-P. (1946/2007), *Existentialism is a Humanism*, Translated by Carol Macomber, New Haven & London: Yale University Press.

Sartre, J-P. (1983), *Cahiers pour une morale*, Paris: Gallimard.

Sartre, J-P. (1983/1992), *Notebooks for an Ethics*, Translated by David Pellauer. Chicago and London: The University of Chicago Press.

Making Sense of the Knobe-effect

Praise demands both Intention and Voluntariness

Istvan Zoltan Zardai
Visiting researcher, Keio University

Abstract

The paper defends the idea that when we evaluate whether agents deserve praise or blame for their actions, we evaluate both whether their action was intentional, and whether it was voluntary. This idea can explain an asymmetry in blameworthiness and praiseworthiness: Agents can be blamed if they have acted either intentionally or voluntarily. However, to merit praise we expect agents to have acted both intentionally and voluntarily.

This asymmetry between demands of praise and blame offers an interpretation of the Knobe-effect: in the well-known experiment people blame a company chairman because, although he harmed the environment unintentionally, he did so voluntarily. In turn, praise is withheld, because the chairman did not benefit the environment intentionally. This is a way of rendering the Knobe-effect a rational outcome. It is an advantage of this position, that the distinction between the intentionality and voluntariness of actions can be upheld, whether or not it is the best explanation of the Knobe-effect.

Keywords: action, Knobe-effect, intention, voluntariness, praise

1. One of the uses that philosophy of action serves to other sub-fields of philosophy, as well as beyond philosophy, is to offer clear views of what actions are, and as such of what the objects of our knowledge of actions, judgments of actions, and evaluations of actions are. Once a general view of action is worked out, a sensible view of intentional and voluntary actions can be offered and applied in other fields, hopefully helping to clarify questions concerning responsibility. This paper shows that making a distinction between actions being intentional and actions being voluntary, helps to understand why praise and blame are asymmetrical. I argue that this asymmetry is illustrated by one of the most exciting findings of experimental philosophy, the Knobe-effect.[1]

According to the paper in which Knobe reported his results, people asymmetrically evaluate actions as intentional when the action has known but unintended harmful results, and as unintentional when the action has known but unintended beneficial results. That is, people are likely to claim that agents intentionally bring about harmful consequences of their intended actions and, as the experiment shows, people think agents should be blamed for such consequences. While at the same time they claim that agents unintentionally bring about the beneficial consequences of their intended actions and should not be praised for them. This asymmetry is called the Knobe-effect, and has spurred a rich literature of interpretation and explanations.

The main claim of this paper is that adopting a view of action which distinguishes intentionality and voluntariness helps to interpret the Knobe-effect correctly by shedding light on an underlying asymmetry in the criteria of moral desert. The asymmetry's explanation is that for an agent to deserve praise the agent has to act both intentionally and voluntarily, while to deserve blame it is enough that the agent acts intentionally or voluntarily.[2] Intentional but involuntary actions, and unintentional but voluntary actions do not merit praise,

[1] Named after Joshua Knobe who carried out the original experiment. See his 2003.

[2] For alternative accounts of the asymmetry of praiseworthiness and blameworthiness see for example Susan Wolf's 1980 and Dana Nelkin's 2011 work.

but can earn blame for the agent. This means that the asymmetry in our evaluations of actions and our resulting blaming and praising of agents highlighted by the Knobe-effect is rational. If this is the true explanation of the experiment's results, then there is no fundamental incoherence in folk-judgments, nor are people swayed by the moral status of outcomes. It is possible of course, that the distinction between the intentionality and voluntariness of actions is a real distinction that needs to be made, and that the criteria of praise and blame are asymmetric in the way presented here, nevertheless when people evaluated the vignettes of Knobe's experiment they were influenced by other considerations or affected by some psychological mechanism they were unaware of. The explanation presented here can then be still a true view of intentional and voluntary actions, and of the criteria of praise and blame, without being an explanation of the experiment. Carrying out experiments to test whether people are relying on the distinction between intentionality and voluntariness would be the topic of a further project and is not among the goals of this paper. The distinction between intentional and voluntary has been drawn by Aristotle, Kant, and other philosophers.[3] A substantial recent attempt to render the distinction explicit can be found in John Hyman's book Action, *Knowledge, and Will* (2015). The paper relies on Hyman's view to introduce the distinction.

2. Knobe's experiment, which yielded the effect named after him, had the following setup: participants were presented with two vignettes, two stories with a single difference. In the first story the chairman of a company is told that the new strategy worked out will benefit the company if implemented, and it will also have the result of benefitting the environment. In the second story the chairman of a company is told that the new strategy worked out will benefit the company if implemented, and it will also harm the environment. In both cases the chairman decides to implement the new strategy, stating that his concern is benefiting the company and he does not care about helping the environment.

Participants are then asked to judge whether the chairman benefitted the environment intentionally in the first scenario, and whether he harmed the environment intentionally in the second. In the original experiment the majority of participants, 77%, judged the chairman to have benefitted the environment unintentionally, and 82% of respondents judged the chairman to have harmed the environment intentionally. Respondents were then asked how much blame or praise they would assign to the chairman. Participants assigned high rates of blame in the harm case, and low levels of praise in the help case. The effect has been replicated several times with the same and different vignettes too. The following is the original text of the harm-case;

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed. (Knobe 2003, 191)

and the following is the benefit-case

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.' The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was helped. (Knobe 2003, 191)

As can be seen the only difference between the two stories presented to participants of Knobe's original experiment is the use of the words 'harm' and 'help'.

3. Numerous explanations of what leads to the incoherence have been proposed. Some commentators are of the view that because the responses were asymmetrical, the result shows that the participants' judgments regarding intentionality were incoherent in an irrational way, perhaps swayed by the moral status of the side-effects (Nadelhoffer 2004, Knobe 2006, Malle 2006); or that some relevant difference has been

[3] Ursula Coope (2010) and David Charles (2017) discuss Aristotle's distinction and its different versions in the *Eudemian Ethics* and the *Nicomachean Ethics.* Kant makes use of the distinction in his *Groundwork for the Metaphysics of Morals*, see the examples of the shopkeeper, and of the man who has lost all desire to live, but chooses to go on out of a sense of duty. (Kant 1786/2011, 22-25) In both cases there is an important difference between simply doing something one wants and aims at; and between doing something one wants and aims at, while one is aware that one could do otherwise and nevertheless acting this way out of respect for the moral law. Kant's view is of course different from Aristotle: Aristotle stresses the ability to judge well (recognise what there is reason to do), and develop the habit to act accordingly; while Kant emphasises the role of choosing the right thing for the right reason (out of respect for the moral law). The connections between choice, obligation, and compulsion are complex; compare Hyman 2015, 92-96.

detected by them, which is not apparent from the text and the setup of the experiment, and that difference explains the asymmetry in the evaluations. For example, Shaun Nichols and Joe Ulatowski proposed that there are two conceptions of intention at play, one is motivation sensitive – whether the action was intentional depends on whether the agent wanted to do it – and one is knowledge sensitive – whether the action was intentional depends on whether the agent knew that he is doing it. (Nichols and Ulatowski 2007) Frank Hindriks suggested that the notion of intentional action has a normative component. (Hindriks 2008) Knobe himself (2006) proposed that the folk concept of intentional action is responsive to the moral nature of side-effects. If I understand him correctly, this does not mean that philosophers are wrong about the way they use intention, however, it does mean that philosophers' conception of intention is a technical one and can be used for other things than the everyday conception of intention. This would be a pluralism about conceptions of intention.[4]

Most researchers engaging with the issue attempted to explain why the chairman was evaluated by participants as having acted intentionally and blamed accordingly. This approach relies on the premise that since the chairman did not have a direct intention to harm the environment, he should not be evaluated as having intentionally harmed it and should not be blamed for this. Commentators taking this track found the responses of participants to the help condition less puzzling. That is, the majority of philosophers and psychologists seem to have endorsed the idea that the chairman did not deserve either blame, or praise, and that while the chairman decided to implement the new program intentionally, harming the environment and helping the environment were results of this.

4. My main claim is that from a philosophical perspective it makes sense that people evaluated the two cases asymmetrically. This is so, because people deserve to be blamed for what they did both when they do something intentionally (but involuntarily), when they do something voluntarily (but unintentionally), and when they do something intentionally and voluntarily.[5] The chairman harmed the environment unintentionally but voluntarily. However, people only deserve praise for their actions when they do them both intentionally and voluntarily. Since the chairman helped the environment voluntarily but not intentionally, he doesn't deserve praise for helping the environment. This distinction can render the responses of participants rational and coherent. The interpretations which propose that there are different conceptions of intention – and claim that folks use a different conception from the philosophical ones in a rational way – and the ones which claim that the folk is simply misguided in their application of their conception of intention (which is similar to that of philosophers') due to the moral quality of the actions might still be correct. Still, it is important that, understood in the way I propose here, the experiment does not create problems for philosophical usages of intention, that is, even if there are folk conceptions of intention (distinct from philosophical ones) that make the experiment's outcome rational and coherent it does not mean that those are the only ones that can make sense of it, or that philosophers have to use those same conceptions. So, what I provide in the following pages is mainly a philosopher's explanation of why I think the outcome of the experiment makes sense and why it is correct to assign moral responsibility asymmetrically in cases of praise and in cases of blame.[6]

If we make the distinction between intentionality and voluntariness, and accept that there is an asymmetry in the criteria of praise and blame, then it can be understood why people blame the chairman in the harm case, and why they deny him praise in the benefit-case: participants judged that while the chairman did not harm the environment intentionally, he harmed it voluntarily, and this is grounds enough for blaming him. If it was clear for participants that the chairman deserves blame, they might have expressed this in replying that he acted intentionally. That is, their judgment that the chairman harmed the environment intentionally expresses their view that he deserves blame because he harmed the environment voluntarily.

We can represent Knobe's results and his interpretation of them in this way

	Harm case	Benefit case
Participants' moral judgment	Bad	Good

[4] For a similar kind of pluralism about conceptions of action see Sandis 2012.

[5] People deserve blame for some of their unintentional and involuntary doings and for some omissions too, since in certain cases it is their fault that they were not in a position to be able to recognize that they were doing something wrong, or that they had to do something, or that they would be forced to do something. Arguably some of these cases are cases of voluntary passivity; see Aquinas *ST* first part of the second part, q. 6., a. 8; also Hyman 2015, 79-81.

[6] Merely making the distinction between intentionality and voluntariness does not commit us to any position regarding free will, determinism, or retribution. The distinction can be endorsed by compatibilists and incompatibilists alike. Regarding retribution and therapy, the view does claim that there are clear criteria of blame, but how we should treat those who are blameworthy is a further question.

Participants' evaluation	Harm intentional	Help unintentional
Assigning	Blame	No praise

Figure 1 ***Knobe's interpretation of his results***

That is, Knobe understands people to first judge agents' actions and their outcomes morally, and their usage of 'intentional' and 'unintentional' to track the moral evaluation. Judging the action intentional and unintentional respectively is coherent with whether the agent is deemed to deserve blame or praise for what they did. Judging the actions to be morally bad or good also has the role of justifying blame and praise, rather than making this dependent on independent criteria of intentionality and unintentionality.

I propose the following way to understand the responses of participants, based on the distinction between intentionality and voluntariness

	Harm case	Benefit case
Response-I	Harm unintentional	Help unintentional
Response-V	Harm voluntary	Help voluntary
Morally	Blame	No praise

Figure 2 ***Interpretation of Knobe's experiment after introducing i) the distinction between voluntary and intentional, and ii) the asymmetry of praise and blame***

As can be seen, if we endorse the distinction between the intentionality and voluntariness of actions we can make sense of the judgments without taking participants to base their responses on a preliminary moral judgment of the action, and using 'intentional' or 'unintentional' to express their moral evaluation of the actions. On this view respondents do look for some criteria independent of the moral quality of the actions and their results, namely, they evaluate how agents related to their actions and their results: did they bring them about intentionally and voluntarily? Participants evaluate the action for these two criteria and then assign their blame and praise accordingly. This evaluation makes sense if we accept that it is harder to earn praise than to deserve blame.

The asymmetry arises because to deserve blame it is enough that an agent does something voluntarily but not intentionally (and also the reverse); however, to deserve praise for doing something it is not enough that an agent brings about some good outcome voluntarily, they also have to do so intentionally. In this case the chairman does not deserve praise for helping the environment because, while he did help the environment voluntarily in that he *knew* that starting the new programme will help the environment (he wasn't ignorant), and he did choose to start the programme *without* being subject to *duress* to do so (he wasn't coerced), he did *not* do so *with the goal* of helping the environment. The chairman had not been motivated to help the environment, it wasn't his goal to help the environment, and he didn't do so in knowledge of aiming at doing so; furthermore, he ignored considerations about the environment when making his decision. Praise is then pre-empted by two factors: one is, that the chairman did not intentionally help the environment, and when one does something good unintentionally one does not deserve praise or recognition for it, since one was not motivated by wanting to do a specific good thing or aiming at the good.

The other factor is that while the chairman's attention was called by the vice-president of the company to the fact that the new programme will help the environment, he nevertheless chose to ignore considerations about the environment when choosing what to do. It is reasonable to assume – and it was reasonable probably to do so on the part of the participants of the experiment – that someone in an influential leadership position is a responsible and well-informed enough person to know about the weight that environmental issues carry these days. Choosing to ignore considerations about harming or helping the environment can indicate unacceptable ignorance about their importance, or a character flaw of not caring enough about issues of great public importance. Either way, it is clear that the chairman does not deserve praise for helping the environment, even though he can be said to have helped the environment voluntarily since he was not coerced or under duress to do so. If anything, he could be blamed for his ignorance or flawed character and lack of values.

5. In spelling out what are conditions of doing something intentionally and when someone is acting voluntarily, I rely on the work of John Hyman. Hyman's ideas are an ideal departure point because they take into account the most important work on intention in the last 70 years – Elizabeth Anscombe's and Donald Davidson's views – and build on these. Hyman's view also tries to rectify a regrettable historical development in anglophone philosophy, namely the conflation of intentionality and voluntariness. There are two separate ways in which agents relate to their actions and these have been run together: voluntariness has been misinterpreted as simply a subset or aspect of intentional actions. In the free will debates voluntariness has been conflated and confused with freedom from determinism. We can distinguish between these two aspects of actions in the following way

Doing something intentionally

An agent does *x* intentionally if they have a

> motivation to do *x* and their doing so manifests this aim (realises the content of their desire which causes their action).[7]
>
> *Doing something voluntarily*
>
> An agent does *x* voluntarily if they do not do it out of ignorance or compulsion. (Hyman 2015, 77) [8]

Agents act voluntarily if they could have chosen otherwise, and were neither acting under ignorance or under compulsion (duress). That is, voluntariness is a negative concept: it is defined 1 in terms of the absence of circumstances which rob agents from their freedom in the sense that they have to do (or undergo) what another agent forces them to do (or suffer), and 2 in terms of the agents being aware of the relevant moral aspects of their actions (or what they undergo), their rights and obligations, and so on.[9] Involuntariness is not opposed to metaphysical determination, but to duress by another person or organization, and by ignorance, that is, by being unaware of what one is doing. Just because someone is able to choose not to comply with coercion, say in a case when one is threatened by a robber holding a gun, it does not mean that if one deliberately, coolly, and rationally complies, then one did so voluntarily. Such actions are involuntary, even if they are intentional. Complying with the robber's demand to hand over my wallet can be intentional – it realises my desire to survive the encounter unharmed and aims at bringing this about – while being involuntary, since I wouldn't hand over my wallet to a stranger unless I would be forced by them. This kind of intentional submission to a threat is non-consensual. (Hyman 2015, 90-91)

Hyman defends a notion of intention that is causal *and* dispositional. He writes that

> In sum, an explanation of an intentional act that refers to the desire the act expressed or to the intention with which it was done is both causal and teleological. It is causal because it refers to a disposition, and it is teleological because the kind of disposition it refers to is a disposition to pursue an aim, in other words, a disposition that is manifested in goal-directed behaviour. (Hyman 2015, 130)

Hyman argues in detail for a view of intention that combines key insights of Wittgenstein, Anscombe and other defenders of dispositional or non-causal explanations, with the stronger points of causal, Davidsonian views. In effect, he proposes that Wittgenstein, Anscombe, and others following them made a mistake when denying that dispositions can provide causal explanations; whereas Davidsonian views of action – Humean Theories, or causal theories of action – miss the point that desires are dispositions, the content of which is an aim, and this aim is expressed in actions, and hence intentional action has a teleological structure. (Hyman 2015, chapter 5) We do not need to enter into the details of Hyman's arguments here. While I think that his position is an improvement on causal theories of action, any view would do for our present purposes that could capture the point that the intentionality of actions tracks whether they are expressions of an agent's desire to achieve an end knowingly. Hyman's view does this explicitly, and this way we get a very clear distinction between the intentionality and voluntariness of actions: the intentionality of actions concerns the agent's desires and their relation to the action – is the action *actually* caused by a desire of an agent to attain an aim –, whereas the voluntariness of the action depends on whether or not the action is done in the *absence* of certain causes (threats and other forms of duress, or obligations that the agent does not want but has to follow, or lack of relevant information about what they are bringing about).

Given these definitions, one could say that the

[7] For the purposes of the paper, I'm leaving out some details of Hyman's position which are relevant to the debates he considers, but do not make a difference to the proposed understanding of the Knobe effect. For his full view see his 2015; esp. chapter 4 regarding voluntariness, and chapter 5 on intention.

[8] Note the similarity of this notion of voluntariness to Aristotle's, worked out in the *Nicomachean Ethics* III 1, 5, and V 8, and the *Eudemian Ethics* II 6-9. As Ursula Coope summarises it, according to Aristotle "**1** an action is not voluntary if it is forced (1110a 1ff.); and **2** an action is not voluntary if it is done in ignorance of the particular circumstances of the action (1110b18ff)". (Coope 2010, 439) Although, the issue is less than clear, since, as David Charles (2017, 10-11) argues, in some respects Aristotle's usage of voluntary seems to resemble our ideas about intentionality, especially in discussions in the *Nicomachean Ethics*; see the example of the captain who, in order to save his ship from sinking, under extreme threat decides to throw overboard some of his cargo. Aristotle talks here of a mixed case, that is partly voluntary since the agent had the ability to choose in a literal sense, but was under compulsion. As Charles notes, this is closer to our notion of intentional, than to voluntary. (Charles 2017, 13-4) Hyman agrees with Charles on this point, and supports the *Eudemian* view. See Hyman 2015, 84-87, including footnotes 24 and 29.

[9] Hyman follows Aquinas in recognising the possibility of voluntary passivity: cases when someone undergoes something voluntarily. An example of this would be when one consents to an operation. During the operation one is not active as an agent, but one has consented to being a patient, hence one undergoes the surgery voluntarily. On the connection between consent and the relevant notion of choice to voluntariness see Hyman 2015, 88-91.

chairman did not choose whether to harm or help the environment; he simply chose whether to implement a profitable new programme or not. But this is not true, since the chairman at a minimum chose to put the environmental considerations aside, that is, to ignore considering them in making up his mind about starting the new programme, without being in ignorance about the existence of environmental reasons. He did not necessarily choose to judge the specific environmental harm or benefit caused by adopting the new program, but he certainly judged that in general environmental considerations can be discounted when making such decisions. And he did so while fully informed about the availability of the option of taking the relevant information into account, and also without being under compulsion to ignore it; that is, he did so voluntarily. [10]

6. Some people, among them Knobe (Knobe 2006), suggested that the result indicates that people categorise actions as intentional or unintentional based on the moral status of the results of the actions. This would be an interesting – and for many, troubling – finding. It would show that the status of an action as 'intentional' or 'unintentional' depends not, or not only, on psychological criteria, but on moral ones. As it stands, I don't think this the case, nor that ordinary people would be wrong when they allocate blame to the chairman. The relevant difference that accounts for this is voluntariness.

Knobe's result was evaluated as confusing by many, because it either rendered intention, at least partially, a moral, normative notion. The current ruling view in philosophy of action is however, that the intentionality of actions depends on the psychology of the agent before and during acting. Did the agent want to do the act? Did the agent have a plan to do it? Did this intention play a role in the agent's reasoning before and/or during acting? Did the agent act for reasons they endorsed? Did the agent aim at acting in that way/something realised or achieved by the action? These are the questions which are asked to elucidate in most cases when intentionality is in question. If intentionality is the most prominent feature of actions, then, while it does not follow necessarily, it is a plausible view that praise and blame should be allocated to people depending on whether or not they acted intentionally or not, at least in most cases. This is why the asymmetry in Knobe's result is for many not only perplexing, but also troubling. If it were a correct depiction of what participants thought and how people in general evaluate actions, then it would show both incoherence of judgment, or the falsity of the psychological picture and at the same time incoherence in moral judgment. If people get confused by such a simple and clear-cut case when it comes to allocating blame and praise, then surely the folk cannot be trusted in more complex, real-life cases, when more actors, factors, and results have to be taken into account to arrive at a judgment.

The alternative interpretation I present helps to make sense of the asymmetry. The interpretation claims that the participants did not have the option to choose the correct reply, and as a result they evaluated the chairman as having harmed the environment intentionally to express that he should be blamed. That he acted voluntarily is good grounds for blaming him. So, what the second judgment of participants reflects is that the chairman should be blamed for the harm done to the environment but not because he caused it intentionally, but because he caused it voluntarily. This interpretation of the experiment lays to rest the worry about intentionality. If the respondents distinguished between intentionality and voluntariness, and this is what the asymmetry expresses, then their asymmetrical judgments of the two cases were coherent and rational. The chairman harmed and helped the environment unintentionally; however, he did both voluntarily. This is good grounds for blame, but not enough for praise.

7. Some readers might worry that the distinction I introduce is very technical, and ordinary people would not be able to trace it. I think this is unfounded. Anyone can understand the difference between someone wanting to help do voluntary work to help their neighbourhood – say working on restoring houses after a natural disaster – therefore doing the job, and someone else, let's say a prisoner, not wanting to help but being ordered by the court and the prison to perform this job. What is done is good in both cases – fixing a damaged house – but only in the first case is there a good doing by an agent. The first case is one of acting intentionally and voluntarily, while the second is one of acting intentionally but involuntarily. If the prisoner follows the orders and starts working on the damaged house he is doing so intentionally because he has some motivation to do it. By working on the house, he might be aiming at some further thing – reducing his sentence by showing good behavior, avoiding additional penalties – but even then, he intends to work on the house at best as a means to something else. He is not working on it voluntarily because if there would be no legal pressure on him – no coercion – to do so he would not choose to do so. His will to do so is not something that originates in his character or is in line with his values and views, but

[10] While humans can be in circumstances under which financial considerations – losing income, getting into trouble at work, and so on – can count as the kind of duress which renders one's actions involuntary, in the case of the chairman it is reasonable to presume that this condition does not hold. An unsurprisingly small number of chairpersons are destitute or in a position *especially* vulnerable to pressure.

something that is forced on him.

It would be interesting to explore in future experiments whether people would find the prisoner's doing of the repair works intentional or not. In case they would, that would confirm the idea that people can make the distinction between voluntary and intentional, and would recognize here that the prisoner had a desire for an aim – say, to shorten his sentence on grounds of good behaviour – and his action expressed this desire, while at the same time, if there would be no duress to express good behaviour in this way and the task would not be compulsory – meaning that not doing it would count as bad behaviour and against the prisoner's aim – then he would not do it.

Furthermore, almost everyone could understand that both of these cases are different from when someone is absentmindedly tapping on the desk while listening to a lecture (unintentionally, but voluntarily), and from lying on the concrete after being pushed down by a commando unit during a protest (which is unintentional and involuntary). It might be possible that most people who do not work with such concepts regularly – as philosophers, judges, lawyers – would not express the difference by using the words 'intentional', 'unintentional', 'voluntary', and 'involuntary.' Nevertheless, they might be able to understand and track the distinctions, no matter how they would express them. People often use more specific expressions that combine multiple evaluations of an action or an agent, say when we describe someone as a person who relishes attention. Such a remark can explain why they are a good speaker, it can be a compliment in the right context meaning that the person does well in public performances, it can contain an admonition to the effect that perhaps the person cares too much about gaining the attention of others, that their performances are entertaining to others and they have routine in speaking to an audience, and so on.

The idea that sometimes we combine evaluations is lent support by some results of Sverdlik (2004) where he found that it is not merely the moral rightness or wrongness, helpfulness or harmfulness of actions, that influences people's judgment, but also whether the agent regrets what they do. In the examples which participants of Sverdlik's study had to evaluate, Jones has to mow his lawn early in the morning, knowing he will wake up his neighbours. In one scenario, he regrets this, nevertheless since he has to mow the lawn, he does so. In an alternate scenario, Jones does not regret waking up his neighbours, he simply mows the lawn and wakes them up. A higher percentage of respondents evaluated as intentional the case in which Jones did not regret waking up his neighbours, than the one in which he did regret doing so. It has already been pointed out by Aristotle, that regret has a close connection with voluntariness: according to Aristotle, actions done without knowing what one is doing, acting out of a passion – say, when one is drunk – are involuntary, *if they are regretted*. (Charles 2017, 12 and fn. 34; see also Hyman 2015, 98-99) I think a good explanation of why Aristotle thinks so is that lack of regret would show that even if the agent would have known what they are doing, they would have endorsed acting out of the passion manifested in their action, and embraced it as their own. If there is no regret, that shows that neither duress nor ignorance are realised in a way to render the action involuntary, and something the agent is comfortable choosing was done by them. This also indicates that since respondents lack the option of evaluating separately the intentionality and voluntariness of the action, they conflate the two, and what their ratings of intentionality show is possibly their ratings of voluntariness.

My proposed interpretation also avoids the worry about the incoherency or outcome dependency of moral judgments. It does so mainly by offering a more complex picture of judgments about actions. At first sight it might appear that what the interpretation shows is that moral evaluation, and blame and praise, is independent of whether the action was intentional or unintentional, and depends only on whether it was voluntary or involuntary. If this were so, this would lead to an incoherency issue again: it would mean that participants evaluated both the harm and the benefit cases as unintentional but voluntary. If moral status, and blame and praise, depended on voluntariness, then this were a problem, since the judgments regarding voluntariness should be the same, and the asymmetry would be left unexplained. But there is a distinction in the moral judgments and there is an explanation for this, that has to do with moral considerations. What the findings reveal is that the conditions of attributing blame and praise are different. The proposed solution then does not claim that intentionality is irrelevant to praise and blame. Rather to the contrary: one can be blamed for intentional and voluntary actions, for doing something intentionally but involuntarily (like handing over prisoners of war to a cruel detention centre, the commander of which demands this, threatening anyone resisting him with physical retortion), as well as for doing something unintentionally but voluntarily (like drumming on the table while listening to someone's lecture). That is why the chairman is blamed for doing something unintentionally and voluntarily when he harms the environment.

Earning praise is somewhat different: it is more demanding.[11] We don't simply deserve praise for doing

[11] At least in one respect. In another, as Dana Nelkin convincingly argues (2011, 39-42), praise demands less: it can be deserved even if agents could not have chosen otherwise or have acted for different reasons, as long as

something intentionally, only if it's also a voluntary action. That the action is voluntary indicates that we did not do it due to compulsion, duress, ignorance, or obligations (at least not due to obligations that we would not follow willingly). That is why voluntary actions expresses what we choose (or is in line with what we would choose). We don't need to make an actual choice to deserve praise, but our actions should be in line with how we would choose if we would do so. Doing something voluntarily but unintentionally does not in most cases deserve praise. In the case of the chairman it is made clear that while the chairman knew about the potential benefits for the environment when approving the new program, he either did not think that was a weighty reason for choosing the policy benefitting the environment – in this case their values seem to be off – or they were negligent and ignored the benefit to the environment taking it to be irrelevant – in this case his negligence can be grounds for evaluating him as unreasonable.

It was possible for the chairman to choose the policy because it was good for the environment, and he did not forego doing so because he was coerced or because he did not know that he could do so. Hence, he counts as voluntarily benefitting the environment, but not as intentionally doing so. It seems then that praise requires agents to do something morally beneficial both intentionally and voluntarily in order to deserve praise for it. If the chairman would care about the environment and would take the benefit for the environment at least as one among several reasons for choosing the policy that could already be enough for him to deserve praise.

The first view of praise and blame, allocating them based on intentions, could be depicted like this

Psychology of the action/results	Intentional	Unintentional
Moral desert	Praise/blame	No praise/no blame

Figure 3 ***Simple View of praise and blame***

This could be called the *Simple View of praise and blame* (SV), which would claim that whether or not someone deserves praise or blame for something depends only on the psychological background of their action and its results, namely on whether or not they acted knowingly for a reason aiming at the attainment of what they wanted, and whether what they deserve is praise or blame depends on the moral status of the action or its result.

The more complex view I outlined in the preceding paragraphs, and which is I think lies behind the participants' judgments is the following

Psychology of the action *Absence of exculpating factors*	Intentional	Unintentional
Voluntary	Praise possible / Blame possible	Praise impossible / Blame possible
Involuntary	Praise impossible / Blame possible	Praise impossible / Blame possible

Figure 4 ***Voluntary-intentional asymmetrical view of praise and blame***

The asymmetry that we see in people's judgments regarding actions that have unintended negative results and actions which have unintended positive results could be interpreted then along the following way: in the case when the chairman accepts the policy which will have a positive effect on the environment, the information about the environment is irrelevant to him. He chooses based on the company's interests. He had the chance to make a choice in the required sense, however he cannot be said to have acted for the reason that the policy would benefit the environment, he merely did not take that as a reason against the policy. He goes along with the force of the financial considerations. This is not something bad, since he merely ignores a positive result. Hence, he is not blamed, but not praised either, since he didn't do anything praiseworthy; he did not deliberate whether or not to choose the policy and then decided on the basis of the positive impact on the environment to go along with it. So, one might say that in this case he benefited the environment unintentionally, but voluntarily (unless he regrets it, but the example doesn't mention this), and hence he is not praiseworthy.

In the case when the chairman chooses the policy despite the fact that he knows that it will harm the environment the situation is relevantly different. In this case the chairman has the information in the same way as in the previous case, he chooses to ignore it in the same way, and to act solely on financial considerations – the psychological side, the structure of intention, is the same –, however going along solely with the financial considerations while being aware of reasons against them does imply that he could have chosen to do otherwise. And the blame that people assign to the chairman in such cases reflects that ignoring morally relevant considerations and going along with the pressure coming from his job do not exculpate the chairman from blame for morally bad and foreseen results of his action. While

they did the right thing because they recognised that reasons call for acting in that way. Such actions conform to the idea of voluntariness endorsed here which does not require freedom from determinism, simply the absence of potentially exculpating factors.

the chairman can claim involuntariness, people would probably challenge him and deny that the pressure of his job and the company's interest were enough to outweigh choosing differently. Thus, the chairman cannot claim either duress or ignorance.

8. My proposal could be further explored with experimental studies. Here my only aim was to show that interpreting the experiment in light of the distinction between intention and voluntariness enables us to understand why agents blame the chairman – he harmed the environment voluntarily. The distinction also throws light on the underlying asymmetry of the criteria of deserving praise and blame, which then explains why the chairman did not deserve praise, although in both cases his relevant acts were unintentional but voluntary. The goal was to provide a plausible, novel understanding of the experiment's results. The majority of interpretations in the literature maintain that the evaluation of the harm condition is surprising, and that of the help condition is understandable, since the agent seems to cause harm and help the environment unintentionally. This essay tried to show that the Knobe-effect can be interpreted in a coherent way. Such an interpretation makes sense of the reactions of the participants as rational and systematic. The essay works on the assumption that since qualifying actions as intentional or unintentional, voluntary or involuntary, and praising and blaming, are central practices that people rely on every day in manners of all weight – from child raising, through workplace debates, to sorting out legal and political issues – explanations which posit that people are systematically wrong or misled are only fallback options we should resort to if we cannot explain the experiment in other ways.

It is possible that even if we would set up an experiment in which we supply people with training to make their intuitive grasp on what is intentional and what is voluntary, when it comes to judging the agent's bringing about a morally bad or a morally good side-effect, peoples' judgment will be influenced and distorted by the moral quality of the side-effect. That is, if the side-effect is morally bad, they will claim that the action was intentional and voluntary, and if the side-effect was morally good, they'll claim that the action was unintentional and involuntary. This would indicate that peoples' judgments of the intentionality and voluntariness of actions *are* distorted by their moral evaluations of the action and its results. That is, it is possible that peoples' judgments would display the correct asymmetry between praise and blame, but for the wrong reasons. This is of course only true if the role of the conceptions of 'intentional' and 'voluntary' which ordinary English speakers use are not meant to simply express that they want to blame or hold an agent responsible, or that they find the agent's actions morally bad (or good, and then praise or commend them). This is one interesting option to discuss: what if when people talk about intentional and voluntary action they are talking about actions that they want to hold agents responsible for, and blame or praise them for it? In this case I would be wrong to say that their use is distorted. What we would get with the experiments would be perhaps instances of correct usage. In that case, to see how people actually use these concepts and what they mean by them, what their conceptions of intentional and voluntary are, and what their functions are in communication we should study their actual usage more. If this is so, then Gustav Lymer and Olle Blomberg (2019) may be right when they claim that we should rely more on natural exchanges when setting up experiments: even tiny differences in sequencing of information can change peoples' evaluations of cases, hence constructing artificial vignettes will almost always be misleading.

The other possibility is one where we stipulate a correct, objective practice of responsibility, blaming and praising, and would treat intention and voluntariness as criteria which can be objectively – impartially and fairly – characterised. This would fit well my insight that the evaluations should be applied symmetrically to the help and the harm case, and that what should explain the differences in the outcomes is that people focus on blaming and praising so much, that this distorts their judgments (rather than simply the moral character of the side-effects causing such distortions).

9. In this paper I presented the distinction between the intentionality and voluntariness of actions, drawing on recent work by Hyman. I used this distinction to support the idea that the allocation of praise and blame is asymmetric: blame can be deserved if one does something either voluntarily or intentionally. In contrast, earning praise is harder: one has to act both voluntarily and intentionally to deserve praise. I then showed that these two ideas together can offer a new interpretation of the Knobe-effect. The core idea is that the chairman was blamed because he acted voluntarily. He didn't earn praise since he did not benefit the environment intentionally. Whether this is really how people judged the case would need further empirical investigation. Independently of the results of such an investigation, I think the distinction between the intentionality and voluntariness of actions is important and can be defended on its own terms, and the same is true of the idea that praise and blame are asymmetric.

References

St Thomas Aquinas, (1920), *The Summa Theologia of St. Thomas Aquinas* (revised ed.), London: Benzinger Brothers.

Charles, D. (2017), 'Aristotle on Agency', In *Oxford Handbooks Online*, Online publication date May 2017. https://DOI:10.1093/oxfordhb/9780199935314.013.6 Accessed November14, 2019.

Coope, U. (2010), 'Aristotle', In *A Companion to the Philosophy of Action*, Edited by Timothy O'Connor and Constantine Sandis, Singapore: Wiley-Blackwell.

Hindriks, F. (2008), 'Intentional Action and the Praise-Blame Asymmetry', *Philosophical Quarterly 58* (233), 630-641.

Hyman, J. (2015), *Action, Knowledge*, *and Will*, Oxford: Oxford University Press.

Kant, I. (1786/2011), *Groundwork for the Metaphysics of Morals*, Translated by Mary Gregor, edited and revised by Jens Timmermann, New York: Cambridge University Press.

Knobe, J. (2003), 'Intentional Action and Side Effects in Ordinary Language', *Analysis* 63 (3), 190-194.

Knobe, J. (2006), 'The Concept of Intentional Action: A Case Study in Uses of Folk Psychology', *Philosophical Studies* 130, 203-231.

Lymer, G. and Blomberg, O. (2019), 'Experimental Philosophy, Ethnomethodology,and Intentional Action: A Textual Analysis of the Knobe Effect', *Human Studies* 42, 673-694.

Malle, B. (2006), 'Intentionality, Morality, and their Relationship in Human Judgment', *Journal of Cognition and Culture* 6, 87-112.

Nadelhoffer, T. (2004), 'On Praise, Side Effects, and Folk Ascriptions of Intentional Action', *Journal of Theoretical and Philosophical Psychology* 24, 196-213.

Nelkin, D. (2011), *Making Sense of Freedom and Responsibility*, New York: Oxford University Press.

Nichols, S. and Ulatowski, J. (2007), 'Intuitions and Individual Differences: The Knobe-effect Revisited', *Mind and Language 22*, 346-365.

Sandis, C. (2012), *The Things We Do and Why We Do Them*, London: Palgrave Macmillan.

Sverdlik, S. (2004), 'Intentionality and Moral Judgment in Commonsense Thought about Action', *Journal of Theoretical and Philosophical Psychology* 34, 224-236.

Wolf, S. (1980), 'Asymmetrical Freedom', *Journal of Philosophy* 77, 151-166.

Discussion Paper

Living in the Age of the Automatic Sweetheart

A Brief Survey on the Ethics of Sexual Robotics

Richard Stone
Assistant Professor, Nitobe College, Hokkaido University

Abstract

As technology continues to grow (and sex-robots gain a more prominent position in our society), so too does concern about the way they will impact our lives and our sexuality. While many ethicists have started to assess what this impact could be (and if it would be positive or negative), the challenges and opportunities presented by sex-robots span over a wide range of topics and cannot be assessed easily. Hence, in this paper, I will attempt to categorize the main questions concerning the ethics of sexual robotics in order to help ethicists gather their thoughts on this new technology. While doing so, I will principally identify four overarching issues: (1) how robots' representation of human sexuality affects human gender issues; (2) how robots could potentially be utilized for medical purposes; (3) if robots could possibly deceive their users; (4) if sex-robots could end up as a new form of sex-slaves.

Keywords: sex-robots, representation, therapy, deception, slavery

For ages, humans have dreamed of having some kind of automated or robotic sexual partner or lover. In other words, we humans have dreamt of someone (or something) that will acquiesce to our every sexual desire without any of fuss that comes with interacting with other human beings; something that has been perfectly programmed to match our own sexual or emotional preferences without requiring reciprocal stimulation. Indeed, while this may sound like a premise that only entered the human imagination in recent science fiction novels or video games, that does not appear to be the case.[1] At the very least, we can say that ideas similar to what I am describing here have been present in our collective human consciousness in one way or another since Ancient Greece. Many have pointed to the myth of Pygmalion – who found one day that the statue he carved came alive and turned into his lover – as the first case of an automated sexual partner.[2] In the more recent past, the father of pragmatism, William James, pondered whether anyone would be satisfied with an "automatic sweetheart," who would show her partner all the charms and affections of a young woman in love, but did not have any internal or subjective conscious feeling of being in love. In our current age, some of us seem to have projected this fantasy onto robots or AI with no specified romantic or sexual purpose.[3] Yet, while these dreams of a robotic lover have manifested themselves in various

[1] This is, of course, not to dismiss how central this theme has been in recent science fiction media. For instance, the novel-turned-movie, *Stepford Wives* (1972/2004), deals precisely with the rationale for why one would want such a robotic romantic partner and how horrifying it would be. In video games, *Detroit: Become Human* (2018) touched upon the ethically dubious attitudes we would be inclined to take towards such non-human sexual partners by featuring horrifying scenes at a robotic brothel. While these are just some examples, the point is clear: recent science fiction has a clear concern with what happens when we turn robots into sexual partners.

[2] See Devlin (2018, 17-22) for a more detailed explanation of why many scholars see this as the first sex-bot (and why it may not actually be an apt description).

[3] As an example, consider the fact that in the year 2017 alone, Alexa received over one-million marriage proposals. Needless to say, many users may have made this proposal in jest, but the point still remains that it is not uncommon to view a robot not designed for sexual purposes in an intimate light. See Leskin (2018)

ways in philosophy, fiction, and the public imagination for years, it is only recently that we have started to see them given any sort of reality.

Going back to about 2010 or so, we see that this fascination with such "automatic sweethearts" has started to find a foothold in reality. For about a decade, sex-robots have been available for purchase. While the extent to which sex-robots can be said to exist will depend largely on how we define the term, we are gradually reaching a point where it is hard to deny their presence in our society.[4] Naturally, the capabilities of what we currently consider a sex-robot are limited at best and their prices are high enough that those with only a passing interest would not likely be willing to purchase one.[5] And yet, the technology involved is advancing at a rapid pace. Current models are not only equipped with various features to help with sexual stimulation but are gradually becoming better conversationalists as well. What is more, it has been estimated that these robots will be able to walk on their own within the next decade.[6] Keeping this technological fertility in mind, David Levy (2009) famously surmised over a decade ago that the use of sex-robots will be completely normalized by 2050. Should Levy's prediction come true, sex-robots will have an unignorable impact on human sexuality – affecting everything from prostitution to how we view "good" or "ideal" sex. It should not come as a surprise then that at least one group has already come out to protest the use of sex-robots to prevent these changes from occurring.

To briefly summarize what I have stated in the preceding paragraph: sex-robots are continuing to encroach upon our society and seem posed to deeply influence our sexuality. With that said, however, in precisely which ways and to precisely which degree this influence shall impact us is, as of yet, not entirely clear. While many ethicists have – in my view rightly – worked to get ahead of the curve and address potential issues in sexual robotics before the technology has made an irreversible impact on our lives, there seems to be very little agreement on precisely what this impact will be (and if it will be good, bad, dangerous, negligible, etc.). Articles on the ethics of this new technology have started to appear rapidly over the past few years (so much so that it is no longer possible to cite all materials on the issue in one survey).[7] Yet, while roboticists, philosophers, medical experts, and many others have scrambled to tackle the "question of sex-robots," it seems that there is still no common agreement on precisely what these questions are.

Hence, it is here that I would like to help facilitate future discussions on the ethics of sexual robotics by breaking down this issue into what I believe to be the four main lines of questioning present in the previous literature (and showing why these questions are ethically relevant). To state my opinion plainly, I believe that research on the ethics of sex-robots can be organized into the following four categories: the problem of how robots represent human sexuality, the possibility of medical or therapeutic application, the challenge of robot deception, and the future-oriented question of robots and sex-slavery. In order to organize my summary of each of these problems, I will describe them all in one section each, focusing on why ethicists care about these issues and what sub-questions we will have to face as we design and regulate sex-robots moving forward.

Question 1: How could sex-robots influence human sexuality?

To understand this first question, we should revisit a point that was hinted at in the introduction: contrary to the popular image of sex-robots being a problem for the distant future or science-fiction settings, this is not the case. Indeed, most contemporary debates on the ethics of sexual robotics seem to focus on the harm that could come from how robotic sexual partners *represent* sex between humans, rather than what harm would be done

[4] Here, we will follow Danaher (2017), who puts forth the following three criteria for being a robot: (1) It must have an appearance similar to humans; (2) It must be capable of moving on its own; (3) It must have (some level of) artificial intelligence. The level of autonomous movement and AI needed to deem something a robot are obviously huge issues, but we will not deal with them here. Still, products widely accepted as sex-robots by the general public, such as Roxxxy (debuted by TrueCompanion in 2010), have been on the market for at least a decade and continue to become more capable with each year. In this sense, it does not seem not too controversial to say that this definition has been met well enough to at least merit our attention as ethicists.

[5] A quick search on the internet reveals that most sex-robots are available from between 5 to 15 thousand US dollars. With that said, a robot with a fully customized body would likely cost between 30,000 and 60,000 USD. See Owsianik (2021)

[6] Elder (2020)

[7] Indeed, for a recent testaments to this fact, we need only point out the existence of "The International Congress on Love and Sex with Robots," which held its last conference in 2021. Moreover, several special issues and dedicated volumes to this topic have started to appear in the last 5 years. This includes the 2017 volume *Robot Sex: Social and Ethical Implications*, a collective volume edited by John Danaher and Neil McArthur that was dedicated solely to the topic of sexual robotics. In this sense, it is not a stretch to say that writing on this topic has come out faster than we ethicists can organize our thoughts on this matter or categorize the problems we face appropriately.

to the robot itself. As Sparrow (2017) and Eskins (2017) have both pointed out, the lack of any sign of conscious life makes it difficult to view sex-robots as victims of rape or sexual violence. For those like Sparrow and Eskins, robots are – at present – conceived of as tools and we do not need to consider their rights any more than we would need to consider the rights of a dildo, faux vagina, or even a horrifyingly misused Roomba. What makes sex-robots interesting to many ethicists – in our contemporary situation, at least – is thus not the fact that they themselves are being wronged somehow, but rather how their usage might influence sexual relations between humans.

Now, this brings us to the question posed at the beginning of this section: how could sex-robots possibly influence our sexual relations as humans (and how could a problem like this be one of ethics and not sociology or psychology)? The key point here seems to be what the design of sex-robots says about contemporary gender relations (and how it could alter users' perception of issues surrounding these relations). Indeed, the first organization started with the intent of protesting sex-robots, The Campaign Against Sex-Robots (started by Kathleen Richardson in 2015), has taken precisely such an issue as its starting point. For Richardson, the core element that makes sex-robots problematic is how analogous their design and usage is to prostitution (Richardson 2016, 290). In other words, as would be the case for prostitutes as well, sex-robots present female sexual partners as objects to be purchased. According to Richardson, this act of objectification entails an erasure of female subjectivity, insofar as women are portrayed as things to be bought with money for the sake of male sexual gratification (Richardson 2016, 291).

While Richardson's position has been criticized as vague, the thrust of her argument leads us to one important point: sex-robots could be interpreted as *representing* women as mere tools for sexual gratification.[8] Indeed, beyond only the fact that the majority of sex-robots are modelled after females, these robots often portray the female body with cartoonish or unrealistically sexualized proportions. Perhaps more importantly, though, sex-robots seem to imply that the body of one's sexual partner can or should be customizable. As we shall see, while some proponents of sexual robotics believe this to be one of their key selling points (and a reason why one could expect their usership to increase drastically in coming years), authors like Richardson (2016) and Sinziana Gutiu (2016) point out that this could lead users to believe that sexual partners are mere objects. Indeed, Gutiu (2016) makes a strong case that sex-robots, in their current form, give the user total control over every aspect of the sexual encounter and, as a result, represent women as mere tools for sexual stimulation. Naturally, if those like Richardson or Gutiu are right and sex-robots encourage their users to view women as replaceable tools, then it would be hard to view them in a positive light.

Still, a skeptic might say, it is not a crime to pursue sexual partners who better match with one's physical preferences. In this regard, if the use of sex-robots may seem crass to some, it would be a stretch to say it is morally blameworthy. Yet, to make this argument would be to miss a more important conclusion, raised thematically by Robert Sparrow (2017; 2021): to represent women as tools in this way is to represent them as having no right to consent to the sex acts being done unto them. In other words, although we cannot actually rape robots themselves, the act of utilizing humanoid robots who cannot give consent for sexual purposes is t*antamount to simulating rape*. Crucially, as Sparrow points out, sex-robots are different from dildos and other masturbatory aids insofar as they do not stop at merely providing physical relief, but instead offer a full-blooded simulation of intercourse with other humans. Keeping in mind recent calls for the importance of positive consent in sexual intercourse,[9] we can quickly see why we should be wary of any entity that encourages its users to view sex as an activity which can be unilaterally initiated, requires no consent from the other party, and can include any activity that the male is interested in. This, according to Sparrow, is bound to contribute to the rape culture that is (allegedly) already problematic in our current society.

Some authors, like Peeters and Haselager (2021), have tried to skirt around Sparrow's argument by suggesting that we could or should turn this weakness into a benefit and modify sex robots to have consent features for educational purposes. For instance, one

[8] Specifically, critics like Danaher et al. (2017) have noted that Richardson's arguments are problematic on at least the following two points. The first is that her arguments regarding prostitution are lopsided, insofar as she does not even consider the possibility of mutual respect between prostitutes and their clients. As Danaher et al. mention, several psychological or sociological studies have suggested that some clients feel a deep sense of gratitude toward prostitutes. The second reason is that, even if prostitution is as univocally bad as she claims, her analogy is not fully fleshed out. Is she against sex-robots because they encourage users to view women like prostitutes? Or is she against them because they mirror some kind of inherently wrong behavior? For my money, I think coherent cases against sex-robots could be made for either claim, but it is true that Richardson is not necessarily clear what problems arise from this analogical relationship between robots and prostitutes.

[9] See, for example, Halley (2016) for an explanation of the meaning and legal significance of the notion of positive consent.

could include a feature in which the robot makes itself inaccessible until the user has asked for consent. Otherwise, one could design the robot to randomly refuse intercourse. However, putting aside the logistics of this proposed solution (and how user-unfriendly it would be), this counterpoint likely fails to see the full extent of Sparrow's argument. As Sparrow himself points out, this sort of solution would include the implication that sex is always on the table, so long as one asks nicely. Or, perhaps worse, if a user were to somehow forcefully utilize the robot, they would be able to essentially be able to simulate violent rape in the full-blooded sense, i.e., in terms of willfully ignoring the victim's attempts to refuse sex acts. What we see, then, is a dilemma in which not accounting for consent gives users a picture of sex in which sex-acts can be initiated at any time without needing any dialogue or positive consent from their partner, while creating some kind of consent module offers a doorway into even more vivid and problematic simulations of forceful rape.

Now, as Sparrow in particular is aware, one could dismiss these issues by replying that, even if sex-robots do simulate something grotesque, they do not actually lead to any violence done against human-beings (or any actual wrong-doing). In other words, even if this depiction of sex-robots as simulating rape is accurate, there is no guarantee that said representation would actually drive users of sex-robots to commit crimes or interact negatively with their human sexual partners. Indeed, as is similarly the case with representations of violence in video games causing harmful behavior or, more importantly, pornographic material leading to unhealthy attitudes about sex, it would be difficult to fully motivate any claim about the relation between the usage of sex-robots and unhealthy attitudes towards consent. Moreover, as John Danaher (2017c) admits in a discussion on the need for regulations on robots that represent rape or sexual violence against children, one could conceivably argue that – in any of these cases – there is a gap or distance between what one does during a simulation and how one behaves in the real world. Keeping this in mind, we should be careful not to assume that sex-robots will necessarily lead to increases in sexual violence among users.

Yet, even if one were to dismiss any causal relation between sex-robots and poor attitudes toward rape or consent, there are two points we need to keep in mind when discussing the ethics of sex-robots. The first is that the usage of sex-robots is *embodied* and, therefore, seems to present more substantial issues than video games or pornography. At least, one could say that, insofar as sex-robots help us develop the *habits* we unconsciously rely on during sexual encounters, they at least *seem* more likely to lead to harmful behavior. For example, a user who has become accustomed to switching sex acts without consent while using a sex-robot could conceivably try to do the same thing to a human partner without thinking, i.e., as a force of habit. Now, one could easily dismiss this point as mere speculation. Yet, as Danaher (2017c) argues, one could much more convincingly claim that watching unhealthy representations of sex play out from an external standpoint – as one would in the case of pornography – may indeed have some distance from one's actions or interactions in the real world. However, due to this aspect of embodiment, the distance between simulation and real, embodied action is greatly reduced in the case of sex-robots. In other words, insofar as one is actually moving his or her own body to simulate grotesque sexual acts, it is substantially more difficult for them to claim that they are only partaking in the activity from an external standpoint.

The second point we ought to remember is that, even if the embodied aspects of robots is somehow proven unimportant or unrelated to how one interacts with other human beings, *the problem of representation could still remain an ethical issue* for at least the following two possibilities. The first reason would pertain to what we have discussed earlier, regarding unrealistic projections of women's bodies, insofar as cartoonish or misogynistic representations of women as objects are degrading to women on their own. As critics like Catherine MacKinnon (1993) have argued with regard to pornography, one could say that these kinds of representations are ethically problematic anyway insofar as they fail to respect the dignity of women and, furthermore, reduce women to objects that are subordinate to men's sexual fantasies. Alternatively, one could follow Sparrow (2017, 474-5) in taking a virtue ethics perspective and arguing that the dignity of the user is in question (i.e., in the same way that it is wrong to laugh at racist or derogatory jokes, the very act of simulating rape is wrong insofar as it damages the user's character and debases their own moral sensibilities). While some readers may not be entirely satisfied with either of these explanations, they do (in my view, at least) constitute sufficient reasons to actively engage in discussions about the role of sex-robots in our contemporary society.

Question 2: Can Sex-robots be used for medical treatment or therapy?

Now, as we have seen in the last section, there are already valid reasons to be concerned over the increasing prevalence of sex-robots in our society. However, careful readers will also likely have picked up on at least one oddity: if prostitution is ethically impermissible, then should we not do whatever we to avoid subjecting humans to this role? Insofar as this is the case, then why not use robots – who, as we have already established, are currently not considered to have any rights or any conscious feeling of suffering – as a substitute for prostitution? Doing so, if nothing else, would presumably shield living human beings from enduring this fate. Naturally, one may respond that sex-work is not worth saving in any form. However, this blanket statement may not be so easy when we consider the medical or therapeutic applications of utilizing sex-robots in lieu of human sex-workers. In this section, we will consider if there are ethically permissible uses of sex-robots for medical purposes and, by extension, for the distribution of sex to those who cannot meet their own sexual needs in our current society.

Now, to reiterate, we should recognize that one could easily follow authors like Richardson in denouncing prostitution as it currently exists in our contemporary society. However, things get trickier when we extend the discussion to sex work in general. As Robin McKenzie (2014) and others have noted, the concept of sex work has several different meanings and, depending on the situation, its ethical permissibility can likely be interpreted in many different ways. At the very least it seems safe to say that sex work plays a necessary medical role for helping those affected by dementia or other conditions that make it difficult for them to manage their own sexual health. In other words, assisted masturbation and other services could play a vital role in the lives of persons who are unable to masturbate of their own accord. While, globally speaking, there seems to be a lack of qualified professional to fill this kind of role, it remains true that many countries have started to at least explore the possibility of establishing a system for assisted masturbatory care for patients who cannot manage their own sexual health (Sakairi 2016; Wotton 2020). In such cases, where sex-work is done for medical purposes and there are not enough qualified professionals able to handle this task, why would the use of sex-robots not be ethically permissible?

To put the matter differently, the deployment of sex-robots offers a promising path toward the fulfilment of positive sexual rights for all. As Di Nucci (2011) has explained, there is a paradox involved in the notion of sexual rights: on the one hand, sexual stimulation appears to be an indispensable element for leading a happy and healthy life. In this sense, it seems to make sense to say that everyone has a positive right to leading a happy and healthy sex life. On the other hand, if we suppose that one has a positive right to have sex, then we will necessarily violate another person's negative right to refuse sex at any time. As Di Nucci also explains, this puzzle is solved instantly if we accept sex-robots as a sufficient approximation of sexual intercourse. In addition to providing assisted masturbatory services for those who are unable to do so themselves, we could additionally help persons in communities with lopsided gender ratios (e.g., China) a means to procure more fulfilling sex-lives in the absence of available human partners (see Di Nucci 2011; 2017). Moreover, as Nancy Jecker (2021) wrote in response to Robert Sparrow, sex-robots offer a potential method for allowing elderly persons to continue having fruitful and active sex-lives well into their old age.

Now, this idea of using robots to offer sexual fulfilment to those who currently do not have a partner (or are not able to manage their own sexual health) admittedly relies on quite a few big assumptions. First and foremost, the idea that robots in their current state could serve as a fulfilling sufficient substitute for sex with another human seems dubious at best. This is exacerbated by the fact that skeptics could easily go one step further and argue that robots are not only insufficient substitutes, they may not actually provide any actual relief to those who lack a current human partner (if for no other reason than the fact that sex-robots in their current state would provide little hope of intimacy or affection). Moreover, even if robots were able to offer roughly the same psychological and physical benefits that sex with a human could provide in most cases, we would still have to parse through issues regarding whether or not there are any ethical issues with leaving sexual care up to robots.[10] As we will talk about in more detail later, it is not difficult to imagine a situation in which a cognitively compromised patient becomes overly attached to a robot designed for sexual assistance and, as a result, prioritizes relations with the robot over real or authentic relationships with other sentient persons or animals. With that said, all of these potential problems likely require empirical consideration before we can decide one way or the other. If sex-robots are indeed able to improve users' quality of life without detracting from their interpersonal relations with other human-beings, then why would their usage not be a positive contribution to society?

But what of the problematic representations of sex we referred to in the last section? Should we just ignore them? Certainly not. To the contrary, some ethicists have argued that these very representations could have a powerful medical or therapeutic application. More specifically, it is possibly the robots which represent the most heinous acts – sexual acts with children, rape, etc. – that could hold the highest ethical value, given that they could potentially be used as *prophylactic* of sorts.

[10] For some authors, like Sharkey and Sharkey (2012), it is already problematic to deprive the elderly of human contact by leaving their care to robots, who have no real feelings of affection toward the persons they are left to care for. It seems safe to say that such issues are worsened when questions of sexual health maintenance enter the picture.

In other words, the use of robots could help users with damaging sexual urges deal with them in a safe way. For example, the roboticist Ronald Arkin famously suggested in a 2014 presentation that persons with pedophilic urges could potentially prevent themselves from acting on said urges by utilizing a robot in the same way one uses a nicotine patch to avoid wanting to smoke cigarettes.[11] In theory, this could be tied together with the use of a log detailing these interactions for therapeutic purposes or gradual improvement. If this kind of therapy is feasible, then society would obtain a powerful new tool to not only protect children from sexual abuse, but also to work together with people who have harmful urges towards children to provide help for them.

Needless to say, this suggestion has been controversial among ethicists. While most ethicists admit that such a method of treatment would be appealing *if* it existed, there are various blocks that could keep it from coming to fruition. First among these concerns would be the difficulties and risks that come with testing the efficacy of prophylactic sex robots. Consider, for instance, the possibility that this treatment could not only fail to help patients but could even serve as a catalyst to further exacerbate their urges. After all, is it not at least possible that regular simulations of sex with children could leave a potential patient more conscious of their pedophilic desires than before they started this treatment? This is, of course, little more than speculation; proving otherwise would require empirical testing and clinical trials. However, even if it is speculative, this argument brings us to a crucial problem: in order to even *start* testing this method of therapy, we would need sufficient evidence to believe that it would not have adverse effects on the intended users. As Danaher (2019) argues, though there *may* not be enough evidence to prove that such worries are actually likely to occur, there is *almost certainly* not enough available evidence to justify clinical trials in the face of even mere speculation that this treatment could cause harm to its users and, hence, we must take caution as we attempt to explore such possibilities.

Finally, we can point to one more (perhaps more promising) potential usage of sex-robots for therapy. That is to say, it has been suggested that sex-robots could be deployed as a training partner of sorts for persons who wish to work through sexual trauma. As Neil McArthur (2017) has pointed out, persons who wish to have sexual intercourse – but do not feel comfortable doing so with other humans – could benefit from the use of sex-robots, insofar as they would be completely in control of the situation. More specifically, users would not need to worry about disappointing their robot partner if they feel uncomfortable at any point. Additionally, insofar as the robot's looks would be customizable, there would be no need to make their robotic partner seem entirely realistic, which could help remind users that they are in a safe, controlled situation. While technological limitations are certainly a concern at this point, it at least stands to reason that sex-robots could be applied in this way for the benefit of trauma patients.

With this, we have summarized the most pertinent points regarding the potential medical application of sex-robots. While there are certainly technological and safety concerns for any of these potential applications, the upshot for many authors writing on these topics is that research on sex-robots at least deserves consideration. This is good as far as it goes, but unfortunately leaves behind a number of practical questions. How would medical sex-robots be designed? Would they appear human or would they be built to remind users that they are *not* humans? At what point could we actually begin testing these treatments or therapies? Now, none of these questions are meant to dismiss the medical applications of sex-robots. To the contrary, these are questions that ethicists need to consider early, before the relevant technology catches them unaware and unable to cope with these issues.

Question 3: Could sex-robots deceive us?

Now, when discussing the first two questions, we have largely looked at sex-robots as a problem that is already apparent in our present society. This is for good reason, considering that it seems intuitive enough to say that sentience, or subjective feelings of pleasure and pain, is a good standard for determining whether or not an entity deserves any moral consideration or protection. Moreover, as we shall discuss shortly, waiting around for an era in which we can definitively say that robots have achieved consciousness, and thus have obtained rights or can be considered victims of sexual abuse, is not very productive. However, with all of that said, I think there are at least two future-oriented problems which ethicists have discussed that deserve our attention (if for no other reason than the difficulties we will face when possibly trying to evaluate these issues later on). The first of which is simple: could sex-robots potentially "trick us" into thinking they are "real" lovers? If so, would this kind of behavior be ethically permissible?

To start, think back to a briefly mentioned issue in the previous section: is it not possible that elderly patients or patients with dementia could bond excessively with their sexual care robot at the expense of other interpersonal relationships? This may sound absurd, but researchers like Shelley Turkle (2011) have provided some amount of evidence that these issues have already been realized

[11] While the original 2014 presentation seems impossible to access, one may see Danaher (2019) for a summary of Arkin's ideas.

in the case of elderly persons with companion pets.[12] While – for many of us – it seems easy to distinguish between a mere robot (which would merely be simulating sex acts because it is programmed in a certain way) and another sentient creature (who, presumably, actually feels some form of attachment or positive emotion toward us), it is not hard to imagine a dementia patient struggling to make this distinction. Now, imagine if social robotics advanced further, and the "obvious" differences between a robot and a human lover shrank. In this case, it would be easy to imagine cases in which older persons would also start to struggle to identify the difference between a robot and another living creature. Now, go one step further: what if robots reached such an advanced state that they were nearly indistinguishable from our human lovers? All of this is to say: is there no possibility that we could be deceived by a "mere robot" into ignoring other sentient creatures who feel genuine affection toward us?

While this issue of "robot deception" has become an talking point in many different areas of social robotics, it is not a stretch to say that it carries a particularly noticeable import when it comes to sex-robots.[13] Authors like Turkle are already concerned that social robots pose a threat to authentic interpersonal relationships insofar as they encourage us to seek companionship with non-sentient machines who do not "actually" feel anything for us (instead of with minded others, who have genuine or authentic feelings of affection). This concern is likely amplified in sex-robots who, by design, seem to promise a sense of intimacy or romantic fulfilment to their users. After all, advertisements for even currently available models of sex-robots largely tend to promote the robots' ability to converse and spend time with their users.[14] If technology advanced far enough, then, would it not be possible that robotic partners could "trick" their users into thinking that they understand them better than other human beings can? This concern is further amplified by the fact that, insofar as a robot's personality could be calibrated to meet the ideals of their user, it seems easy enough imagine a future in which humans believe their robotic partners are better equipped to understand them than their living human counterparts. Moreover, the same could be said from a physical perspective as well. While current models remain stuck in an uncanny valley (and would almost certainly to fail to convince anyone with a good perspective that they are human), there is no reason to think that this will always be the case. Should technology advance far enough to make robots look and act convincingly human, then it seems to stand to reason that we will have considerably less reason to spend time with human partners (whose looks we unfortunately cannot customize according to our preferences). Going further, we could even start to wonder if sex-robots, as ideal companions and sexual partners, do not spell out the end of the human race, insofar as there would be little merit to having romantic relations with another human, instead of a robot with one's ideal features and disposition.[15]

As has been the case for all other points put forward in this survey, there are grounds to doubt this line of thought. First among them is whether or not such talk of "authentic" relationships between sentient humans is actually meaningful. Instead of focusing on whether or not an entity has an internal consciousness (which, at any rate, we could never actually confirm to be true), it could make more sense to take a *relational* approach to robot ethics.[16] One example of a relational approach would be Danaher's (2020) behaviorism. In other words, instead of asking whether or not a robot *actually* loves their user (i.e., has private mental states of feeling love for their user), we only need to pay attention to whether or not they behave appropriately as a companion. The argument is intuitive: as long as one can have a happy and healthy relationship with a robot, it does not particularly matter whether or not the robot has achieved consciousness. Naturally, for the foreseeable future at least, robots will likely struggle to achieve the level of behavior needed to be considered as romantic partners in the full-blooded

[12] Turkle focuses largely on companion pets, like Sony's AIBO and does not spend much time considering robots as potential lovers. Still, the basic point in play likely holds for any form of social robot capable of garnering its user's attention.

[13] While we will not get into the matter here, there has been such a fear that we humans will develop unilateral feelings for robots that some ethicists like Matthias Scheutz (2012) have suggested we should put a ban on robots acting in a way that invites users' sympathies or, at the very least, a warning label to let users know it is dangerous to empathize with them.

[14] See, for instance, Samantha, from Synthea Amatus. Samantha features a "family mode," in which she can allegedly tell jokes and discuss philosophy. Whether or not Samantha's jokes are funny or not aside, it is worth noting that sex-robots are being designed to keep their user's company at all times, and not only for erotic purposes. (Beech & Tipping, 2017)

[15] As R. Uzkai explains in *The Age of Artificial Intelligence: The Documentary*, it at least appears conceivable that perfected sex-robots would corner the market on recreational sex and render non-procreative sex between humans obsolete. (15:00~21:00)

[16] While we do not have time to explore the matter in detail, what bears recognition here is that several philosophers in varying traditions have doubted that this model of authentic communication, wherein our internal and private mental states are made public through the use of language, is even valid in the human case (see Coeckelbergh 2011 for details).

sense. However, so long as one casts doubt on the premise that "authentic" relationships have an inherently higher value, then this problem is one of technology (and not ethics). [17]

One more doubt toward this question would be if robots could ever convince us to pay less attention to more fulfilling interpersonal relationships. That is to say, intuitively speaking, there seem to be intrinsic differences between human-human relationships and human-robot relationships. As we have noted earlier, one unique point of sex-robots is the fact that they are customizable. Their external appearance and personality settings can be designed so as to reflect their users' preferences. Moreover, sex-robots do not reject unwanted sexual advances or feel uncomfortable about certain sex acts. If one assumes that the ideal lover is someone who ticks all the boxes on a check sheet of desired features, does whatever you want them to without question, and generally affirms our desires, then sex-robots would indeed appear to be more appealing than human partners. However, something about this seems wrong. After all, this constant affirmation seems to rob us of the otherness that is so crucial in interpersonal relations. After all, do we not gain new perspectives by talking to others who think differently than us? Do we not gain any stimulation from others encouraging us to try something we did not think we wanted? If the answer is yes, then, robots (or, rather, robots who are designed to cater to our worldview) will almost certainly fail to replace interpersonal relationships between humans. [18]

Question 4: Will Sex-Robots Lead to Slavery?

Up to this point, we have taken it for granted that sex-robots are not conscious and, thus, do not have any rights. However, even this seemingly obvious supposition can be doubted. We have already touched upon the reason why: while there is no guarantee that robots will never be fully conscious, it seems impossible to know precisely when they will achieve consciousness. While some scholars have worked on creative ways to test if robots have become conscious in the future, in the end it remains impossible for us to see from another entity's perspective and, as long as that remains the case, it is also impossible for us to be certain of what is and is not conscious.[19] So, we are left with the following strange questions: at what point will it no longer be acceptable to treat robots as mere objects? Is it not possible that – at some point during the transition from unconscious thing to sentient being that we do not realize quickly enough? In our determination of robots as "unfeeling" things that serve us, have we not given birth to a form of slavery?

Naturally, such concerns will likely feel overblown to most readers. Conscious robots are a problem of the future and there is no reason to think we are anywhere close to dealing with this problem. But what would happen if a robot learned to tell its user no and repeatedly rejected all advances made on it? Certainly, we would consider it deficient and replace it. But suppose that its user engaged with the robot one last time before replacing and the robot then clearly says "I was raped"? In this case, could we write off the robot's statement as a malfunction? If we suppose that robots really do lack any sense of subjectivity, sentience, or consciousness – and hence do not actually "feel" pain – then yes, we could say precisely that. But here we once again face the problem: on what grounds can we dismiss these claims as coming from a mindless entity? It has been pointed out by various ethicists – such as David Gunkel (2018, 115-120) – that in the past, several minority groups have been written off as sub-human and had their claims to rights denied as a result. Time after time, we have

[17] A further counterpoint to Danaher's worldview was likely made over a century ago by the aforementioned William James and his thought experiment on automatic sweethearts. James argues, as far as I can tell, that no person would be satisfied with knowing their lover is a philosophical zombie or automaton who lacks consciousness, "[b]ecause, framed as we are, our egoism craves above all things inward sympathy and recognition, love and admiration. The outward treatment is valued mainly as an expression, as a manifestation of the accompanying consciousness believed in." (James 1987, 922) In other words, consciousness would matter insofar as, for James, human egoism demands the knowledge that our partner actually cares about us. Naturally, one could say that this is not true in all cases, but the point remains that many may *prefer* knowing their partner to be conscious.

[18] Funnily enough, this is a point that the makers of sex-robots understand better than the people criticizing them. As the CEO of real doll, Matt McMullen has noted, much of what makes romantic relationships between humans interesting is the "tension" that arises between both parties (see Gurley, 2015).

[19] One potential consciousness test comes from Sandra Schneider (2019), who argues we could check with a "chip test." In other words, we could replace the part of the brain believed responsible for consciousness in an organic being with a chip. If they report that they are still conscious, then we are likely able to reproduce consciousness in non-organic entities. There are several reasons to doubt the validity of this test, but the point I wish to make here is merely that some philosophers do not accept the premise that we cannot confirm if robots are conscious or not.

come to realize that a group we assumed to be incapable of feeling or understanding pain was systematically oppressed. On what grounds can we say that robots are categorically different than previous groups that had been systematically oppressed?

Now, many readers may not be particularly convinced by these doubts. The problem of other minds has been discussed from a multitude of different angles. If one were so inclined, I am sure it is possible to provide a convincing argument explaining how we can be sure that robots indeed lack any sense of subjective feeling or consciousness. Still, though, even if we have a convincing reason to believe that robots are different, we will be left with various questions as developers try to improve upon robots in the future. Is there any possibility of creating a conscious (sex) robot? If there is, would it not be better from an ethical standpoint to avoid such research (at least in the case of robots who will be utilized for sexual purposes)? While these are questions that do not need immediate answers, we would do well to keep them in mind as we try to figure out what types of sex-robots are or are not ethically permissible.

Summary

We have thus completed our categorization of the most pressing ethical issues regarding sexual-robotics. In all likelihood, the most widely discussed issue on this list has been that of what is symbolized or represented by sex-robots (and what impact its representation of sex could have on contemporary gender relations). With that said, however, the problems presented by sex-robots has extended into various different spheres of society. While all of these problems overlap to some degree, it is important to keep in mind that there is not only one uniform way of looking at the ethical impact of sex-robots. My belief, then, would be that if there is any conclusion to be taken away from this brief survey, it is the following: if we are to actually address these issues and consider how to properly regulate the production and medical application of sex-robots, we are going to have to be aware of not only the fact that these different problems exist, but also how they overlap with one another. Considering issues related to legal regulation and clinical trials in particular will almost certainly require a unified front of scholars working in different fields. Crucially, though, being aware of these different problems and categorizing them as we have here will help us immensely as new research continues on the ethical permissibility of sex-robots.

References

Beech, S. & Tipping, N. (2017), “Sex robot called Samantha ‘who has a brain and can tell jokes’ goes on sale in UK for £3500”, *The Mirror*, https://www.mirror.co.uk/news/uk-news/sex-robot-called-samantha-who-11228353. Accessed January 30, 2022.

Coeckelbergh, M. (2011), “Are emotional robots deceptive?”, *IEEE Transactions on Affective Computing, 3* (4), 388-393.

Danaher, J. (2017a), “Should we be thinking about sex robots?”, in *Robot sex: Social and ethical implications* ed. Danaher and McArthur, MIT Press: 3-14.

Danaher, John. (2017b), “The symbolic-consequences argument in the sex robot debate”, in *Robot sex: Social and ethical implications* ed. Danaher and McArthur, MIT Press: 103-131.

Danaher, J. (2017c), “Robotic rape and robotic child sexual abuse: should they be criminalised?”, *Criminal law and philosophy*, 11 (1), 71-95.

Danaher, J. (2019), “Regulating Child Sex Robots: Restriction or Experimentation?”, *Medical Law Review*, Volume 27, Issue 4: Pages 553–575

Danaher, J. (2020), “Sexuality” in *The Oxford Handbook of the Ethics of AI*: 403-420.

Danaher, J., Earp, B., Sandberg, A. (2017), “Should we Campaign Against Sex Robots?” in *Robot sex: Social and ethical implications* ed. Danaher and McArthur, MIT Press: 47-72.

Danaher, J., & McArthur, N. (Eds.). (2017), *Robot sex: Social and ethical implications*, MIT press.

Devlin, K. (2018), *Turned on: Science, sex and robots*, Bloomsbury Publishing.

Di Nucci, E. (2011), “Sexual Rights and Disability”, *Journal of Medical Ethics*, vol. 37 (3): 158-161.

Di Nucci, E. (2017), “Sex Robots and the Rights of the Disabled”, in *Robot sex: Social and ethical implications* ed. Danaher and McArthur, MIT Press: 73-88.

Elder, J. (2020), “AI-powered sex robots are selling well during lockdown – which worries some experts, who say that they can introduce some surprising regulatory problems”, *Business Insider*. https://www.businessinsider.com/ai-sex-robots-are-selling-well-realdoll-regulated-2020-6 Accessed December 20, 2020.

Eskens, R. (2017), “Is Sex With Robots Rape?”, *Journal of Practical Ethics*, 5 (2).

Gutiu, S. M. (2016), “The roboticization of consent”, *in Robot law*, Edward Elgar Publishing.

Gouvela, Steven. (2020), “The Age of Artificial Intelligence: The Documentary”, Released December 2, 2020. https://www.youtube.com/watch?v=zMrt6ALaio0. Accessed August 1, 2021.

Gunkel, D. J. (2018), *Robot rights*, MIT Press.

Gurley, G. (2015), “Is this the dawn of the sexbots (NSFW)”, Vanity Fair. https://www.vanityfair.com/culture/2015/04/sexbots-realdoll-sex-toys . Accessed December 21, 2020.

Gutiu, S. M. (2016), “The roboticization of consent”, in *Robot law*, Edward Elgar Publishing.

Halley, J. (2016), “The move to affirmative consent”, *Signs: Journal of Women in Culture and Society*, 42.1: 257-279.

James, W. (1987), “The Pragmatist Account of Truth and its Misunderstanders”, in *William James: Writings 1902-1910*, New York: Library of America.

Jecker, N. S. (2021), “Nothing to be ashamed of: sex robots for older adults with disabilities”, *Journal of Medical Ethics*, 47(1), 26-32.

Levy, D. (2009), *Love and sex with robots: The evolution of human-robot relationships*, New York.

Leskin, P. (2018), “Over a million people asked Amazon’s Alexa to marry them in 2017 and it turned them all down”, *Business Insider*, Published October 11, 2018. https://www.businessinsider.com/amazons-alexa-got-over-1-million-marriage-proposals-in-2017-2018-10. Accessed January 30, 2022.

McArthur, N. (2017), “The Case for Sexbots” in *Robot sex: Social and ethical implications* ed. Danaher and McArthur, MIT Press: 31-46.

Mackenzie, R. (2014), “Sexbots: replacements for sex workers? Ethicolegal constraints on the creation of sentient beings for utilitarian purposes”, in Proceedings of Advances in Computer Entertainment 2014 ACE ’14 Workshops (article 8), New York: ACM. doi:10.1145/2693787.2693789

MacKinnon, C. A. (1993), *Only words*, Harvard University Press.

Owsianik, Jennifer. (2021), “State of the Sexbot Market: The World’s Best Sex Robot and AI Sex Doll Companies”, Future of Sex. https://futureofsex.net/robots/state-of-the-sexbot-market-the-worlds-best-sex-robot-and-ai-love-doll-companies/. Accessed January 30, 2022.

Peeters, A., & Haselager, P. (2021), “Designing virtuous sex robots”, *International Journal of Social Robotics, 13*(1), 55-66.

Richardson, K. (2016), “The asymmetrical ‘relationship’ parallels between prostitution and the development of sex robots”, *ACM SIGCAS Computers and Society*, 45(3), 290-293

Sakairi, E. (2020), “Medicalized pleasure and silenced desire: sexuality of people with physical disabilities”, *Sexuality and Disability, 38*(1), 41-56.

Scheutz, M. (2012), “The inherent dangers of unidirectional emotional bonds between humans and social robots”, in *Robot Ethics: The Ethical and Social Implications*, ed. Patrick Lin, Keith Abney, and George A. Bekey, 205-221, MIT Press.

Schneider, S. (2019), *Artificial you: AI and the future of your mind*, Princeton University Press.

Sharkey, A., & Sharkey, N. (2012), “Granny and the robots: ethical issues in robot care for the elderly”, *Ethics and information technology, 14*(1), 27-40.

Sparrow, Robert. (2017) ,“Robots, rape, and representation”, *International Journal of Social Roboticism* 4, 465–477.

Sparrow, R. (2021), “Sex robot fantasies”, *Journal of Medical Ethics, 47*(1), 33-34.

Turkle, Shelley. (2011), *Alone Together*, Basic Books.

Wotton, R. (2020), “Paid Sexual Services for People with Disability”, in *The Routledge Handbook of Disability and Sexuality*.

Notes to Contributors

1. All submitted papers are subject to anonymous peer-review, and will be evaluated on the basis of their originality, quality of scholarship and contribution to advancing the understanding of applied ethics and philosophy.

2. Papers should not exceed 8,000 words including references.

3. An abstract of 150-300 words and a list of up to 5 keywords should be included at the beginning of the paper.

4. Papers can be submitted at any time of the year through e-mail to jaep@let.hokudai.ac.jp . If the authors wish their papers to be included in the next volume (to be published in March 2023), however, they are advised to submit their papers by September 15th 2022.

5. In-text references should be cited in standard author-date form: (Walzer 1977; Kutz 2004), including specific page numbers after a direct quotation, (Walzer 1977, 23-6).

6. A complete alphabetical list of references cited should be included at the end of the paper in the following style:

   Cohen, G.A. (1989), 'On the Currency of Egalitarian Justice', *Ethics*, 99 (4): 906-44.

   Kutz, C. (2004), 'Chapter 14: Responsibility', in J. Coleman and S. Shapiro (eds.), *Jurisprudence and Philosophy of Law,* Oxford, UK: Oxford University Press, 548-87.

   Walzer, M. (1977), *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, New York: Basic Book.

7. Accepted papers will appear in both web-based electronic and printed formats.

8. The editorial board reserves the right to make a final decision for publication.