Ethics Inf Technol (2013) 15:235–238
DOI 10.1007/s10676-012-9305-y

BOOK REVIEW

David J. Gunkel: The machine question: critical perspectives on AI, robots, and ethics

MIT Press, Cambridge, MA, 2012, 272 pp, ISBN-10: 0-262-01743-1, ISBN-13: 978-0-262-01743-5

Mark Coeckelbergh

Published online: 3 November 2012

Are machines worthy of moral consideration? Can they be included in the moral community? What is the moral status of machines? Can they be moral agents, or (at least) moral patients? Can we create "moral machines"? In the past decade, there has been a growing interest in these questions, especially within the fields of what some call "robot ethics" and others "machine ethics". David Gunkel's *The Machine Question*, however, differs from much of the existing literature since it does not aim to offer answers to these particular questions. Instead of arguing for a particular position in the debate, the author attends to the question itself. What is the question we are asking if we ask about the "moral considerability" of machines? How is the problem framed? What does this frame reveal and what does it exclude?

Gunkel begins his book by remarking that "the machine question" is new: for most of Western intellectual history, technology has been defined in an instrumental way (p. 6). Even philosophical work on the moral considerability of animals is relatively recent. But whereas today many people accept that (some) animals deserve our moral consideration, machines remain the excluded 'other': 'the other that remains outside and marginalized by contemporary philosophy's recent concern for an interest in others.' (p. 5) Throughout the book, Gunkel describes many of the mechanisms of this exclusion. For example, he draws our attention to Descartes's "beast-machine" and to how the moral line between humans and things is drawn by the words "who" and "what". But he also shows that current attempts to end this exclusion are highly problematic.

In the first chapter the author shows many of the complications with arguments for including machines as moral agents. For example, he shows not only that certain concepts such as consciousness (p. 55) and personhood (p. 90) are problematic—this is generally acknowledged by philosophers across the spectrum—but also that there are epistemological problems involved with ascribing these properties to other entities. In particular, as Dennett and Derrida point out, participants in the debate make a 'leap from some externally observable phenomenon to a presumption (whether negative or positive) concerning internal operations'; such an inference is unfounded (p. 64). In Chapter 2, some similar complications are discussed. For example, it turns out that the criterion "can they suffer?" is problematic. What is suffering? Is it the same as feeling pain? Does it require consciousness? Again the epistemological question is raised: 'If animals (or machines) have an inner mental life, how would we ever know it?' (p. 117).

Furthermore, Gunkel agrees with Birch (1993) that the very work of demarcation, every attempt of drawing a line between those who are part of the club and those who are outsiders—the very effort to apply criteria of inclusion/exclusion—is an act of violence. The very way the 'Machine Question' is asked legitimates the domination and exploitation of others (p. 30). In particular, Gunkel shows that "the machine" plays a key role in this process. Showing how Descartes divided human beings from the animal–machine, Gunkel argues that 'the machine is not just one kind of excluded other; it is the very mechanism of the exclusion of the other.' (p. 128) Indeed, as Gunkel puts it:

> 'whenever a philosophy endeavors to make a decision, to demarcate and draw the line separating "us"

M. Coeckelbergh (✉)
Department of Philosophy, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: m.coeckelbergh@utwente.nl

> from "them," or to differentiate who or what does and who or what does not have a face, it inevitably fabricates machines.' (p. 207)

Thus, "machine" becomes synonymous with that which is situated outside our moral consideration. Can we think otherwise?

In the third chapter Gunkel explores an ethics that 'operates beyond and in excess of the conceptual boundaries defined by the terms "agent" and "patient": it is a 'deconstruction' of the agent-patient opposition itself (p. 8). What does this mean? It does not mean a more inclusive ethics. He shows that both exclusion and inclusion are problematic: they are two sides of the same coin (p. 162). Furthermore, he also dismisses a social-constructivist approach. To say that agency and "patiency" are socially constructed, Gunkel argues, is a mere conceptual inversion, which shakes things up but remains 'within the conceptual field defined and delimited by the agent-patient dialectic' (p. 10).

In order to think otherwise, he uses Levinasian thought. If we apply Levinas's thought that there is first the ethical relationship and then cognition (p. 176), Gunkel argues, we must say that there are not 'first' agents and patients which then have encounters. It is the other way around:

> 'The Other first confronts, calls upon, and interrupts self-involvement and in the process determines the terms and conditions by which the standard roles of moral agent and moral patient come to be articulated and assigned.' (p. 177)

This means that Levinasian thinking 'does not make prior commitments or decisions about who or what will be considered a legitimate moral subject' (p. 179). Following Calerco (2008), Gunkel infers that this opens up the possibility 'that anything might take on a face' (p. 179). 'Anything' must be taken literally here; Gunkel suggests, against Levinas, that not only humans but also animals and indeed *things* such as machines can take on a face. (In which case, I presume, we would no longer *call* them "things".)

One may object that this book does not give us answers, whereas we need answers in ethics. Surprisingly, Gunkel uses Descartes to answer this objection: he rejects Descartes's epistemology all along, but at the end, he endorses Descartes's view concerning the provisional nature of ethics (a point which he could have developed by drawing on the pragmatist philosophical tradition). So we can use ethical guidelines, for sure, but they will always be provisional (p. 212–213).

Gunkel's deconstruction is a *tour de force* that largely succeeds in getting us to 'think otherwise'. His argument that machines have always been the excluded other is convincing, and he shows a way to move outside the conceptual space of most current discussions in machine ethics. It must also be mentioned that Gunkel engages the thinking of both 'analytic' and 'continental' philosophers, which I believe is a virtue. However, I see the following issues that need further attention.

First, Gunkel puts much emphasis on the epistemological problems. But as Torrance (2012) has pointed out in a recent conference discussion, the "other minds" problem which Gunkel uses to criticise 'moral agency' and 'moral patiency' arguments, is a typical Cartesian problem. The 'considerable uncertainty' (p. 88) about inner states of other entities he thinks we are left with is Cartesian. But this is exactly what Gunkel wanted to avoid. He (and many others) wish to move beyond Descartes, but here he uses an argument that still remains within the Cartesian 'game'.

We should further note that on the one hand, Gunkel wants to avoid "anthropocentric" domination (we decide if other beings are members of our club). On the other hand, he thinks that it is up to us to decide:

> 'we, and we alone, are responsible for determining the scope and boundaries of moral responsibility (…). We are, in effect, responsible for deciding who or what is to be included in this "we" and who or what is not.' (p. 215)

Thus, Gunkel presupposes that moral consideration is something that is and should be under human control. In my recent book *Growing Moral Relations* (Coeckelbergh 2012)—which also points to epistemological problems and bears striking similarities to Gunkel's questioning of the question since its critical gesture consists of attending to the conditions of possibility of moral status ascription—I question the 'control' assumption with regard to moral status. How we relate and how we should relate to other entities is not entirely up to us; it depends on a range of linguistic, social, and technological conditions; it is not 'made' but it evolves, it grows. Gunkel might be happy to acknowledge this, but his stress on decision and his quasi-existentialist emphasis on the enormous responsibility we have are likely to divert attention from this aspect: he assumes that we have the power to decide who is in and who is out. Also, Gunkel's suggestion that moral consideration is a matter of human decision seems inconsistent with his own criticisms of this view. For example, he thinks it is problematic that human beings 'not only get to formulate the membership criteria of personhood but also that they nominate themselves as the deciding factor' (p. 66). It turns out that his view is also "anthropocentric", to use Gunkel's own term. In spite of his efforts to do otherwise, Gunkel's proposal risks to reproduce the 'basic power structure of anthropocentric (…) ethics' (p. 126).

A more relational approach, by contrast, would not talk about the status of "the machine" but develop the thought that moral status 'is something that comes to be conferred

and assigned to others in the process of our interactions and relationships with them.' (p. 91) In this view, moral consideration or moral status is not entirely our human decision; it rather grows in the process, in the relations we have with (what become) others.

This point transitions to the next problem. Gunkel's (and Levinas's) abstract 'face' of the 'other' needs to be contextualized and historicized: 'the face' takes shape within these relations and becomes a concrete, flesh-and-blood (or metal-and-wire?) face in the process. Levinas's 'face' becomes ethics itself, it becomes something absolute; this ethics seems to be blind to the link with the particular other. Moreover, my making the other the absolute other, sameness seems to be excluded. The other is never entirely other.

More generally, Gunkel suggests throughout the book that machines have a face (or can have a face), but he never really explains what this means. Of course it would be mistaken, in Gunkel's view, to answer this question by specifying the criteria that a machine must satisfy in order to be considered as having a face. But the reader is left too much in the dark about its meaning. We want more comprehension, and comprehension need not be total or totalizing. Perhaps we need descriptions of phenomena that would count, according to the author, as involving a machine face. We want to know what it means to say that a machine has a 'presence' that 'demands recognition, caring, and shared pain', to use Haraway's words (p. 131). Gunkel does not want to provide answers. But he raises a question, a new question: "Do machines have a face?" or "Can machines have a face?" These new questions themselves need to be questioned. What are we asking when we ask if machines can have a face?

Maybe we need a reflection on machine faces that mirrors Haraway's comments on Derrida's cat. Gunkel quotes Haraway (2008), who speculates that Derrida 'began each morning in that mutually responsive and polite dance' but that this 'embodied mindful encounter did not motivate his philosophy in public' (p. 123). Perhaps reflections on machine faces also need to be fed by our (and the author's) experiences of 'embodied mindful encounters' with machines. This would—literally—give 'body' to the now all too abstract arguments. This is true for 'the face' but also for 'the machine'. What is "the machine"? What would it mean to feel its "presence" or to "respect" it? From a relational point of view, we can only know what these terms mean within relations and on the basis of our experience of these relations. From the perspective of epistemology, when it comes to moral considerability we can do without criteria and tests, but we need our, human experience. What is missing here are descriptions of (imaginary or historical) face-to-face encounters between humans and machines. With Levinas, we should emphasize that the encounter is prior. The theorizing follows. In Gunkel's thinking, the face appears first as a word, as *logos*. The flesh and the wires, the presence and the relation, remain in the background. To the extent that presence is reduced to logos, his thinking remains logo-centric (to use a term discussed by Derrida). Gunkel is right to say that we must question without end 'what respond means' (p. 216). But this is what is still missing in the book: what "respond" means needs to be described and, to the extent that it cannot be described, it needs to be *shown*. Where is experience?

These questions, in turn, raise the following problem: whereas most of us have experience with animals, the same is not true for, say, intelligent robots. If this is the case, how much *sense* does it make to talk about, for example, "the face of the robot"? What *is* our responsibility if we lack the experience? Levinas's reflections were written in response to very concrete World War II experiences. What experiences, exactly, does Gunkel have in mind when he talks about the face of the machine? Computers? Current robots? Future artificially intelligent machines? Whether or not he has in mind science–fiction narratives, we want to know more about the phenomenon, about the experiences and their history. Like Levinas, Gunkel runs the risk of dehistoricizing the face. The concrete, ethically pregnant encounter disappears and is replaced by a more abstract ethics of the face. (At the same time, although he criticizes narratives of inclusion, given the many analogies he draws between the history of animal ethics and machine ethics he seems to uncritically embrace the story that machines will become increasingly intelligent until they will be included in the moral community.)

Considering Levinas's ethics, the more fundamental issue here is: "what kind of vulnerability do machines have?"; and a related question has to do with "what kind of *violence* can be done to them?" If we want to apply Levinas's conception of ethics to machines, we need to be much more specific about what could happen to machines. The discourse on animal ethics, for example, is based on a shared understanding that *there might be a moral problem in the first place* in our dealings with animals, and this shared understanding is based on knowledge and experience in particular, historically and geographically situated encounters with animals. In particular, it is based on knowledge about *violence* against animals: not only the conceptual violence Gunkel, Derrida, and others discuss, but also real violence, the violence that hurts (even if the latter is connected to the former, as Gunkel suggests). It seems that the face of the machine can only appear to us as a face if and when we see that vulnerability and that (possibility of) violence. The future of machine ethics thus depends on the kind of relations we will develop and have with machines. To let their moral considerability depend upon *proof* is highly problematic, as Gunkel's book shows. But whether or not machine faces appear—now or in the

future—depends on what we do to them, whether it is dancing or fighting. This renders machine ethics "anthropocentric" in this, relational sense. However, like in the human and animal case, acknowledging this link between human–machine relations and moral consideration does not necessarily exclude that we extend our moral consideration to machines. On the contrary, the latter seems to presuppose the former.

An additional problem concerns the way Levinasian ethics understands "the social": like Levinas, Gunkel seems to limit the social to I–you relations. The social appears as "others". But social reality cannot exclusively be described in these terms. The 'face of the other' does not account for the full richness of social experience. Moreover, the relation with the particular other is always mediated by culture. Often the other's 'face' does not even appear to us—that is, the other does not appear as an other—because our culture frames the entity we encounter in different terms. This is, arguably, what happens and happened with the machine. Gunkel describes the way of thinking that excludes, but does not put this way of thinking in a social-cultural context. If we think in a certain way, for example if we are somehow stuck in agent–patient thinking, then this is not only 'my' responsibility but 'our' responsibility, and it remains questionable how much control we have over these larger socio-cultural patterns. In this later work Heidegger suggested that the way we think about technology is not something that we can change by an act of will. This may put constraints on 'thinking otherwise' in the more mundane sense of changing our views of machines (the conditions of possibility I discuss in my book can understood as constituting such constraints).

Finally, with Benso (2000) Gunkel dismisses interpreting the technological other in terms of technè, which he understands in terms of art or 'technology and its aberrations' (p. 192). But a different, broader understanding of technè could leave room for other kinds of human–technology relations: relations that make possible engagement, responsibility, and, perhaps, the appearance of the face. If we keep focusing on "the machine" rather than our relations with technology, we might well remain blind to different epistemic and ethical possibilities. A non-instrumentalist understand of technology also implies that individual machines are not necessarily the most appropriate unit of analysis. The face may not end at the border of the machine.

In spite of these problems (and, like any good work, partly *because* of the issues it raises), this book is an original contribution to the field, and it is likely to have wide reverberations. Asking questions about the question enables Gunkel to shed new light on a problem that will remain of central importance not only within machine ethics but also within ethics in general. As Rorty already suggested in the 1990s, perhaps the question of who should be part of our moral community is *the* ethical question. We are already moral beings; the question is not why we should be moral but how far our moral gestures should reach. Or to say it in Gunkel's words: the question is where the face begins and where it ends. Even philosophers who disagree with the author's Levinasian view, might be inspired by his approach to the problem: by his efforts to go beyond 'more of the same' and his insistence that we should not stop questioning, that every morality is provisional.

In addition, Gunkel's careful and remarkably comprehensive review of the literature will also be useful to readers new to the field of machine ethics. But more importantly, this book shows that good, critical philosophical reflection on machines is not only about how we should cope with machines, but also about how we (should) think and what role technology plays (and should play) in this thinking. This book is not 'only' about machine ethics because, paradoxically, it shows that modern ethics has always been a machine ethics. For this reason, the present reviewer hopes that the book's readership will extend to those philosophers who still assume that ethics has nothing to do with machines.

References

Benso, S. (2000). *The face of things: A different side of ethics*. Albany, NY: SUNY Press.

Birch, T. H. (1993). Moral considerability and universal consideration. *Environmental Ethics, 15*, 313–332.

Calerco, M. (2008). *Zoographies: The question of the animal from Heidegger to Derrida*. New York: Columbia University Press.

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. Basingstoke/New York: Palgrave Macmillan.

Haraway, D. J. (2008). *When species meet*. Minneapolis, MN: University of Minnesota Press.

Torrance, S. (2012). The centrality of machine consciousness to machine ethics. Paper presented at the symposium 'The machine question: AI, ethics, and moral responsibility', AISB/IACAP world congress 2012—Alan Turing 2012, 4 July 2012.