Is Act Utilitarianism Self-Effacing?
The Rising Need of Utilitarian Awareness in Indirect Strategies

SUZUKI Makoto
Associate Professor, Nagoya University

Abstract: *Ethical theories can be "self-effacing": as we are less successful in achieving their point when we hold and apply them, they enjoin us to stop doing so. It has been argued that self-effacing theories are problematic. Act consequentialism, for example, act utilitarianism, remains the most prominent target for this charge. Because act utilitarianism is the most familiar version of act consequentialism, this paper focuses on that view, even though the ensuing arguments apply mutatis mutandis to consequentialism in general.*

Act utilitarianism might well be self-effacing when the ways of holding and applying the principles are restricted in a certain manner (say, to deduction), but act utilitarianism itself does not involve such restrictions. Act utilitarianism can be held and applied with smart thinking devices, and through the finding and improvement of these devices, it has become less prone to self-effacement.

One might suspect that it is not necessary for everyone to accept and hold the principles of act utilitarianism in order to find out and improve thinking devices that achieve the act utilitarian purpose. After all, intellectual elites can do that for other people. So act utilitarianism might still endorse partial self-effacement. Such a worry about partial self-effacement will be addressed in view of various contrivances that purport to direct people's decision making and conduct.

As science and education have progressed and spread, act utilitarianism has increasingly ceased to be self-effacing. However, it could be self-effacing in the past. As the verdicts of utilitarianism rely on input from the real world, perhaps this world-dependent result on its self-effacement should be expected. However, this type of historical consideration is rarely found in the literature. The question of self-effacement awaits more thorough scientific investigation, particularly into the trajectory of actual human history.

Ethical theories can be "self-effacing" (Parfit 1984, Sections 9 and 17): as we are less successful in achieving their point when we hold and apply them, they enjoin us

to stop doing so. Many people have argued that self-effacing theories are problematic: they fail to properly guide action, violate the publicity condition,[1] disrupt agents' mental harmony or integrity, and/or require agents to have conflicting moral judgments (Smith 2001). While other ethical theories, for instance virtue ethical theories, are also allegedly self-effacing (Keller 2007), act consequentialism, for example, act utilitarianism (AU), remains the most prominent target for this charge. Though there are various characterizations of self-effacement (Cox 2012, 290), act consequentialism is often supposed to fit all of them.

Like many philosophical debates, arguments about the self-effacement of ethical theories tend to focus on what implications the phenomenon would have, and whether they would be problematic (see, e.g. Parfit 1984, Smith 2001, Cox 2012, and Akiba 2016). However, these questions are preceded by another genuine issue, i.e. whether theories in question are really self-effacing. This paper concerns the latter question for act consequentialism. Because AU is the most familiar version of act consequentialism, this paper will focus on that view, though the ensuing arguments apply mutatis mutandis[2] to consequentialism generally.

It is probably true that, when applied to human beings in the actual world, AU enjoins us to have motives and decision-making procedures other than trying to do what the principle tells us, i.e., to maximize the happiness of those concerned. Utilitarians have long since acknowledged this point (e.g. Mill 1861, Chapter 2). However, for AU to be self-effacing, it must enjoin more, i.e. ensure the removal of 'itself' in the agent's mind. Therefore, we have to ask: is AU really self-effacing?

1. The Terms of the Debate

Though there are various definitions of self-effacing moral theories, I essentially follow Parfit's (1984) classic characterizations:

> A theory is completely self-effacing iff it tells everyone to believe some other theory.
> A theory is partially self-effacing iff it tells some people to believe some other theory.

[1] According to the publicity condition, a moral theory must be such that everyone should accept and acknowledge it to each other (Rawls 1971, 133 and 182).
[2] That is, by replacing reference to welfarism with another value theory.

A theory is self-effacing for a person P iff it tells P to believe some other theory.

As I understand AU, it dictates that an agent ought to maximize the sum (or average) of well-being. In talking about the maximization of well-being, I mean what actually makes the world go best in terms of well-being. I am not going to talk about a theory that dictates that an agent must maximize the expected well-being or utility, i.e. the weighted average of all possible well-being resulting from the action, with the weights being assigned by the expectation degree (i.e. subjective probability) that any particular event will occur. Such a subjective theory is usually not taken to be self-effacing to the extent that its objective counterpart is. The threat of self-effacement looms when an agent might not do what a theory recommends if s/he believes and tries to follow it. It seems that, given that the agent's beliefs determine which options have the best expected utilities, s/he has a good chance of doing what a theory recommends, i.e., taking the option that has the best expected utilities, if s/he believed and tried to follow it. Of course, the subjects might tend to miscalculate the expected utility or spend too much time calculating it. It is still apparent that the epistemic problem is still larger for the objective variant; if even the objective version has resources to address the problem of self-effacement, the subjective variant has better resources. This paper regards objective AU as the main focus of this debate. From now on, when I say "act utilitarianism" (AU), I refer to the objective version.

Whether a theory is self-effacing depends not only on the content of the theory but also on the empirical facts about the agents and their natural and societal environment. If an agent were omniscient, s/he could figure out and perform what maximizes well-being. Therefore, for him or her, AU would not be self-effacing. If everyone found themselves in the world where Descartes' evil demon (Descartes 1641, Meditation 1) dominates, AU, along with many other moral theories, would be completely self-effacing, telling everyone to, say, ignore the principle and do what s/he believes to be harmful. However, these far-fetched possibilities are not the focus of this debate. I assume that agents and environments are actual human beings and actual human conditions.

2. Conditions on Application?

AU might well be self-effacing when the ways of holding and applying the principle is restricted in a certain manner. For example, if it is required that the application of ethical principles like AU be consciously deductive, then, given our favoritism and limited knowledge, we would often have mistaken factual premises and end up producing suboptimal outcomes. For example, suppose a student has a choice between going to see his academic advisor at the time he promised to do so, or not going to see her and instead study for his exam. If he held AU and applied it deductively, he might well reason as follows:

> I should take an option iff it would maximally promote the well-being of those concerned.
> Skipping the appointment and studying for my exam would maximally promote the well-being of those concerned.
> Therefore, I should take the option, i.e., skipping the appointment and studying for the exam.

As a result, the student would skip the appointment and study for the exam. However, it might well turn out that this decision would have suboptimal outcomes. For example, the student would miss the opportunity to get important information on the exam and the defense of his thesis in a timely fashion, the relationship between the student and the advisor would be damaged, and so forth. If he had believed some deontological view that one should always keep a promise no matter what, or the non-welfarist (and hence non-utilitarian) view that keeping a promise is intrinsically important, he might well fulfill the goal of AU better.

For another example, if it is presupposed that the ways of applying AU are those the agents would choose 'naturally', i.e., if they held the principle and were given no instruction on how to apply it, then again, the agents would often produce suboptimal outcomes. Consider, for example, a notorious case where you walk near a shallow pond where a child is drowning. In the 'natural' application of AU, you consider which option would maximally promote the well-being of those concerned. You consider which option exits. It might take some time to conclude that there are three options: leaving the place without doing anything, calling the police, and entering the pond and saving the child. You try to calculate which option has which outcomes with what probabilities, and manage to figure out that you should enter the pond. When you finally entered the pond to save the child, the child has drowned. If you believed a deontological view that you should save a person when you can no

matter what, or a non-welfarist view that a human life is intrinsically and non-comparatively important, you might well have saved the child and achieved the goal of AU better. The time and opportunities that utility calculus costs are often very expensive.

If AU must be held and applied in the above ways, it will often have suboptimal outcomes. AU might support believing another theory and applying it if these cases are frequent and the alternative theory would have a better record.

Perhaps there must remain a few people who hold AU because, if all of us ceased to believe AU, people's beliefs in moral theories would be without an act utilitarian check and they might start acting in a very anti-utilitarian way. This reasoning given by Parfit (1984, Section 17) is convincing, but partial self-effacement is still a real possibility if the way of holding and applying AU is restricted in the above ways.

The question is whether these restrictions must be accepted. You cannot deduce from conceptual analysis that AU must be applied either consciously deductively or 'naturally'. AU is the standard of permissibility, and it does not by itself determine how it should be used in guiding thought and action. Of course, as you will see in the next section, given that believing and applying a theory is itself an action, AU would recommend whatever way of believing and applying the theory, which would maximize the well-being of those concerned. However, the concrete way is determined not conceptually, but by the world we live in. For example, it is not guaranteed that utility calculus is part of the recommended way of applying act utilitarianism.

Some might suppose that the recommended or canonical way of believing and applying a moral theory must be deductive or 'natural'. These people have a particular conception of what it must be like to be guided by a moral theory, for example, that moral principles must be held and applied in the same manner as laws, or that its application must take the form of justification.[3] Such a conception is controversial and not part of AU itself, and its advocates must present some positive argument that AU must satisfy this conception.

3. Believing AU and Applying it with Smart Thinking Devices

[3] Another idea of application is that an application of a theory must be such that an agent cannot fail to achieve the standard or aim set by the theory. I think that in this sense of application, human beings cannot apply not only AU but also its main rivals, deontological theories and virtue theories included.

I have argued that AU can be self-effacing when the ways of holding and applying the principle are restricted in a certain manner. However, often self-effacement will not be optimal provided that we have the options of holding and applying AU with the aid of various decision guides, for example, with well-made secondary rules, simplified decision trees, or computers running utility enhancing programs.[4] These thinking devices can be established through the *after-the-fact* investigation of *tendencies*, rather than the *prospective* utility calculus about each *particular action*. The former is obviously easier than the latter.

As an example of well-made secondary rules, I present three major rules of safety engineering: inherent safety, safety factors, and multiple barriers (e.g. Hansson 2014, Section 4). These rules are the rules of thumb regarding serious risks. Inherent safety dictates that you eliminate a hazard itself rather than trying to reduce risks associated with that hazard. Safety factors dictate that the strength of designed constructions exceed certain numerical factors in order to ensure that they are stronger than the bare minimum expected requirement for their functions. Multiple barriers dictate that each barrier to the relevant risk be independent of its predecessors so that if the first barrier fails, the second is still intact, et cetera. These rules are not introduced as utilitarian second rules, but they can be adopted as such provided that they would increase the well-being — or in this case, reducing its loss — better than any other known alternative guides, including utilitarian calculus. The above set of three rules, or something like it, might well satisfy this condition at this point in time; satisfying them prevents numerous risks from being actualized, and the cases of not satisfying them, for example, that of the Fukushima Daiichi nuclear plant, faced severe outcomes.[5]

As an example of simplified decision trees, I present one from Gigerenzer's book (2007, 174: see Figure 1). This is the tree used for deciding in a timely fashion whether a patient with severe heart pain, which may be a symptom of a heart attack,

[4] Note that I am not arguing for rule utilitarianism (RU). RU holds that an action is permissible iff it complies with the system of rules the application of which would promote well-being more than the application of any other systems of rules. The position in the text does not assert that rules, or any other thinking devices, determine which action is permissible. It still embraces AU, which holds that the standard of permissibility is whether the action itself maximally promotes well-being. Even if an agent uses the thinking devices, they occasionally act sub-optimally, which is not permissible according to AU.

[5] The Fukushima Daiichi nuclear plant put all the electronic sources along the coast side, which is a violation of multiple barriers. A tsunami destroyed all of them, and this is the direct cause of the meltdown.

should be sent to a coronary care unit or to an ordinary nursing bed with an electrocardiographic telemetry (ibid. 168). This "fast and frugal tree" succeeds better than not only specialized doctors' unaided intuitions, but also complex statistical methods, in correctly sending heart attack patients to coronary care units while not sending so many non-heart attack patients there (ibid. 175; see Figure 2). The tree helps to make quick life-and-death decisions.[6]

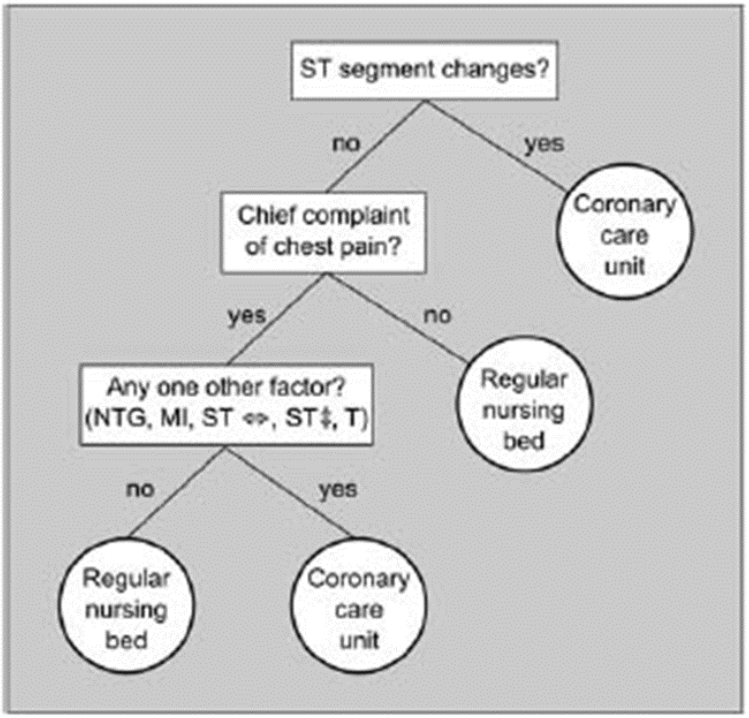

Figure 1: The Fast and Frugal Tree for Deciding the Treatment of a Patient with Heart Pains (Gigerenzer 2007, 174)

[6] As Gigerenzer notes (2007, 176), a full decision tree has so many branches that it is computationally intractable and unsuitable for timely practical decision. What he calls "first and frugal" trees consist of three blocks; first, looking up for factors in order of importance; second, stopping a search if a factor allows it; and third, classifying the object according to this factor (ibid. 175).

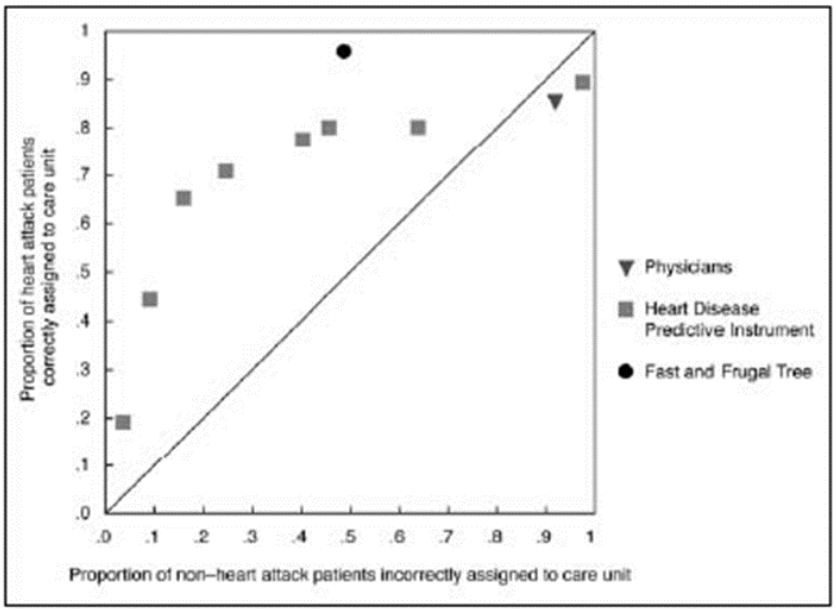

Figure 2: The Proportions of True Positives (Vertical Axis) and of False Positives (Horizontal Axis) (Gigerenzer 2007, 175)

As for a utility enhancing computer-run program, I do not know any that is actually used for this very aim. However, it will resemble the one that is used in determining who receives a particular organ that is provided by a brain dead donor in Japan (Japan Organ Transplant Network 2015, especially 10). Because of the lack of organs, the problem of distribution is severe in Japan. A computer-run program is set up to select the recipients that satisfy certain necessary conditions of compatibility and that maximally satisfy certain additional criteria. By the necessary conditions of compatibility, I mean the conditions that must be satisfied in order for the organ to do any good to the recipient. For example, if the donor has B-type blood, the recipients must have either B-type or O-type blood, because otherwise the transplanted organ would not function in the recipient's body. Given these inputs, the computer will select the recipients. Because the additional conditions reflect not only considerations about efficiency but also those about medical immediacy, order on the waiting list and so forth, this might not exactly fit the description of the utility enhancing computer-run program. However, by modifying the content and priority of the conditions, we can construct a program for implementing AU in so far as our current conditions of knowledge allow. If human beings have to determine who will

receive an organ or who will not, it would be not only too time-consuming to transplant it timely, but also a stressful process because it is potentially a decision of life and death. Additionally, the computer-run program can avoid the influence of doctors' favoritism for their own patients. Using such a computer-run program can be a good guide in the application of AU in such a domain.

All of these thinking devices have appropriate domains of application. If act utilitarians make use of them, they will employ different devices in different domains.

The point is that one can adopt the above decision guides *while holding that AU is true and that these guides are devices to achieve the aim of AU*. Because these guides are not flawless, they might end up producing suboptimal outcomes in certain cases. Thus, the agents who deploy them do not believe that using them would maximize well-being on every occasion. I still take it that this mode, i.e., holding AU and using the thinking devices to achieve its aim, is a way of applying, or of trying to follow, AU. Given our limited knowledge, these methods have better chances of achieving the aim of AU than any other known alternatives, including the use of utility calculus.

Furthermore, AU would not enjoin us to forget itself and take these thinking devices as the ultimate criteria. Doing so would eliminate the chances to improve these devices on the bases of new data and their analysis. Through the observations of "experiments in living" (Mill 1859, Chapter 3) and real controlled experiments, i.e. on the basis of scientific investigation, we can improve these thinking devices; they are first discovered on the basis of experience, so in principle they can be improved upon thorough further empirical investigation. However, in order to judge that they are improved from the act utilitarian viewpoint and to adopt them for this reason, we need to keep our belief in AU. Because we do not naturally have the maximization of well-being as the object of our primary desires, believing in AU as the guiding principle is crucial for our adopting and revising devices-cum-actions from the act utilitarian standpoint.

4. The Possibility of Partial Self-Effacement

At this point you might argue that everyone's holding of AU is not necessary for finding out and improving upon thinking devices to achieve the act utilitarian purpose. Intellectual elites can do that for other people. Therefore, AU might still

endorse partial self-effacement: only these elites should hold AU. This conclusion amounts to a variation of the position that Sidgwick (1907, 489) warily argued for and Williams (1985, 108–109) made famous in his critique under the name of "Government House Utilitarianism".

Sidgwick (1907, 490) argued that in certain situations, people should not know utilitarianism, since the calculation of utility is so difficult that their actions lead to bad results. However, this paper has argued that AU recommends people to use smart thinking devices over utility calculus as the means to achieve the utilitarian end. As far as such devices are available,[7] people generally do not need to make calculations, and their holding of AU does not produce the bad results Sidgwick is concerned about. The above paragraph, however, points out the possibility that as far as smart thinking devices are available, ordinary people do not need to keep AU in mind. Only intellectual elites hold AU and create these devices, and ordinary people merely use them.

There are four reasons why AU would not support such an intellectual division of labor in the actual world. First, if the creation and revision is up to elites, their self- or group-interests might well distort the thinking devices so that their use will promote their interests but have suboptimal outcomes in terms of general welfare. In order to check this bias, ideally every competent agent should hold AU to make sure that the thinking devices tend to promote the well-being of everyone concerned.

Of course, we might not need AU to recognize direct disadvantages against ourselves and our families because we are naturally sensitive to such effects. However, some relevant effects of using a thinking device might be easily overlooked, such as those involving long-term consequences, and outcomes to strangers and those who cannot voice their complaints (including non-human animals). By holding AU, people recognize them as theoretically important and pay attention to these effects from time to time.

Second, there are so many differences in people's abilities and social situations and what makes them better-off, that what works for some people (e.g. intellectual elites) might not work for other people. From the act utilitarian viewpoint, this means that even if a thinking device is usable by some group and tends to promote general welfare in their situations, it might not be usable by us or tend to be rather suboptimal in our situations. Even if technology-dependent thinking devices, which involve computers, the internet, and so forth, are available,

[7] Section six examines what happens where this assumption does not obtain.

some people might do better in their own preferred way; the use of paper maps might work better for you than the most recent computer navigation system. To create feasible and workable thinking devices for diverse populations in different circumstances, their users and those potentially affected by the uses should be able to voice their relevant experiences and expectations. What counts as relevant experiences and expectations depend on the guiding principle, AU. By helping people recognize the principle and through using it to guide public discussion, people will pay attention to things that are relevant to general welfare and provide information that may be more useful in improving upon thinking devices. Thus, in developing the thinking devices for AU, people are generally advised by AU to hold it as the standard.

Third, in the use of thinking devices, we need to take AU as setting the point of use. This need comes from three facts. First, usually, thinking devices can be employed in more than one way. This tendency is strengthened when we try to develop general thinking devices for various people and situations. Often, without recognizing the point of employment which is given by AU, one cannot choose the proper use. Second, because the users are human, there are risks of error in employing the devices. Realizing the point of use will help people to avoid rather obvious versions of these errors, for example, misreading or misinterpreting the instructions for use. If they realize that their way of employment will lead to suboptimal results and this will violate the point set by AU, they might well suspect that something is wrong and come to correct their mistake. Third, even if the thinking devices determine the one definite way of application and we avoid human error, thinking devices will sometimes lead to suboptimal results. This is unavoidable. These devices are made to promote general welfare most efficiently in the long run, and not in each and every occasion. To expect more than this is beyond us, the beings whose prediction abilities are severely limited. This condition is aggravated as these devices are always under development to deal with new circumstances and to become more user friendly, more efficient, and open to wider application. The holding of AU will help people recognize rather obvious cases of suboptimal outcomes and, in those cases, to find quick fixes for application procedures.

Fourth, in order for smart thinking devices to be known and used, education is necessary and some justification is required to put them on the lists of things to be taught. Otherwise, people will not keep them in mind and voluntarily use them. If people are taught AU first, we can provide them with a unified and straightforward

rationale that explains why using such diverse thinking devices are important, which might well be more persuasive than piecemeal and disconnected explanations.

For these four reasons, ordinary people need to keep AU in mind when using thinking devices. AU does *not* support the intellectual division of labor where only intellectual elites hold AU and develop these devices and ordinary people merely use them.

5. Externalization and Partial Self-effacement

However, you might suspect that we do not need education, or at least a rationale for education, to propagate and run thinking devices. We can externalize the thinking devices, e.g. by putting well-made secondary rules into laws or professional codes of ethics, or simplified trees into official guidelines for practitioners. Bentham and his fellow reformers attempted to do this in 19th century Britain. While they tried to institutionalize secondary rules with punishment, economists often attempt to do so with positive as well as negative incentives. For example, some of them recommend that the government provide insurance premium deduction, which makes it a rule that people take out certain insurance.

As Lessig (1999) points out, physical or technical constraints on activities (e.g. locks on doors or firewalls on the Internet) can constrain us as these regulations can. Such "architectures" can also be regarded as externalized directive. A recent, subtle implementation is a nudge, which "is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting fruit at eye level counts as a nudge. Banning junk food does not" (Thaler and Sunstein 2008, 6). In certain domains, by nudging people in certain ways they will promote general welfare more efficiently than coercing them in those directions or leaving them to act as they wish. If these attempts of externalization succeed, people can achieve the aim of AU by following externalized thinking devices. Even the illiterate can be expected to act in the act utilitarian way.

There are differences in the transparency of these contrivances. Laws and official professional codes of ethics are relatively transparent to the intended addressees; they know these public rules and they recognize that they are supposed to follow them. However, incentive structures and architectures, including the

setting of a nudge, often work more covertly. The intended addressees might not recognize what direction these environmental conditions lead them to, or even that they are ways of courting their decision making in a certain direction. They often do not use these devices consciously.

Now let us go back to the issue at hand. The suggestion is that we might not need education, or at least a rationale for education, to propagate and run thinking devices because we can externalize them. People do not have to know the reasons for the legalization and codification of secondary rules and simplified decision trees, because these laws and guidelines come with their own punitive structures. Putting an incentive structure and nudging in place might not require education at all, as people's decision making is not affected by the recognition of their hidden direction. Since knowing the doctrine does not affect people's decision making and action, AU does not require people to hold AU, or so argued.

However, we need to keep AU in order to develop or check externalizing thinking devices in the proper way. Here "we" not only include those in charge of externalization, but also the people who use the externalized devices and who are affected by their use. In the previous section I mentioned that losing sight of the standard has four disadvantages: there will be a greater risk of the devices' being created only for the sake of the developers of the device; it will be difficult to develop the device to accommodate the diversities and complexities of people and environments; as people will not recognize the point of running the devices, the results might severely violate the end of AU in certain cases; and lastly, education on thinking devices might well become less convincing. The argument in the above paragraph concerns only the last disadvantage. The first three cautionary remarks apply to the externalization of these devices as well.

Let us think, for example, about incentive structure and nudging. These interventions are often so subtle that they have a considerable risk of being used for special interests and of having their invisible and often long-term effects overlooked. Keeping AU in mind, people will have a greater chance of looking into these points in view of general welfare. And because incentive structures and nudges do not work in the proper way in certain situations, it will be beneficial from an act utilitarian view that people can check whether it promotes general welfare in the particular situations they encounter. Though some might think a priori that an incentive for an action always tends to induce that action, this is not true, and we need to examine it on a case-by-case basis. For example, Thaler and Sunstein (2008, 232) suggest a program for reducing second pregnancies of teenage girls: city

governments pay girls who have already had a baby a dollar for each day they are not pregnant. As Gigerenzer performed a literature study, he found only one randomized trial by Stevens-Simons et al. (1997), which reported that giving a dollar a day did not change the rate of second pregnancies. As this example illustrates, humans and their societies are so complex that it is hard to know how a particular incentive or nudge will work out for the intended group in the intended situation. Sometimes incentives even hamper the intended action by crowding out people's social preferences (see Bowles 2016, especially Chapter 3 and references therein). From the act utilitarian standpoint, it will be beneficial if people hold the doctrine and can examine whether the incentive or nudge actually promotes general welfare. Such an examination will involve subtle consequences, including people's motive structure (ibid. Chapter 6).

Note also that externalizing a thinking device has its own cost. For example, enforcing a criminal law requires police, a justice system, jails, and the money to support them, not to mention the decrease in welfare from the restriction of freedom. Given these costs, AU will expect people to pay attention to and check whether these externalizations will really promote general welfare most effectively.[8] In order to do so, an act utilitarian viewpoint is necessary.

6. Individual and Social Conditions, and the Prospective Effacement of Self-Effacement

If my argument is on the right track, the resulting style of application is not to recognize AU and then use it with empirical premises to deduce what one should do. Nor is it to be unaware of AU and use a thinking device instead. Rather, it is to be aware of AU and implement a thinking device or its externalization as the best feasible way to achieve its end. The direct awareness of the principle, and its indirect application through such a device will be the strategy for success in view of AU. AU is not even partially self-effacing.

Or is it? Until now I have supposed two things: the availability of such thinking devices as explained in section three, and the abilities that are sufficient for using and checking thinking devices or their externalizations to achieve the end of AU. We need to admit that these conditions are not universally satisfied.

[8] Nudging allows choice and costs little in terms of personnel and money, but often leaves people uninformed (Gigerenzer et al. 2009).

The availability of such devices or their externalizations is not universal. Because well-made secondary rules are "general conclusions from the experience of human life" (Mill 1861, Chapter 2), they could not be established without empirical investigations. Even after they are established, they might not reach the general population. For example, many people do not know the three maxims of safety engineering. Furthermore, technology-dependent thinking devices, which involve language, paper, printing, computers, the Internet, and so forth, are simply unavailable if these technologies are absent, unaffordable, or presented without instructions and interfaces that are understandable to them. In the past, these devices were not found at all. Then, trying to apply AU might have involved a 'natural' application which would have included utility calculus. The result might have been suboptimal and even worse than believing some other theory.

As for the abilities to use thinking devices to achieve the end of AU, these devices are often made for normal human beings. If some people do not have the capacities of normal adult human beings, their use of these devices might have suboptimal results. For example, if some people cannot enter inputs into the relevant computer properly, the computer-run program will not return the desired results. For this task at this point, literacy and the ability to understand certain basic instructions are necessary. Perhaps future technological development will make these devices less demanding, but currently some people cannot use these devices to bring about desirable outcomes.

Thus, perhaps AU has been self-effacing for the ill-informed and the illiterate, and even for many normal adults who applied it in the past. Certain restrictions on its application, which is discussed in section two, is forced by their personal or social conditions; and, in believing and trying to follow AU, they might end up facing results worse than believing another ethical view available to them at that time.[9]

However, given the development of natural and social sciences as well as the improvement of linguistic and other literacies, it might well cease to be so, as human beings gain more accurate information and have access to better decision guides in applying the principle of AU. Moreover, if the thinking devices for AU are successfully externalized, more people can hold and apply AU successfully with the help of these externalized devices. Informing people of the principle has become

[9] This partly depends on what other moral views have been available. Note that if these theories are unavailable to an agent at that time, holding them will not be his or her option, so it will not be recommended by AU.

generally conducive to the aim of AU in rendering thinking devices and their externalization more efficient. AU recommends that we work on science and education that have generally improved the conditions of human welfare. Therefore, act utilitarians will seek a world that is increasingly developed in terms of science and education, which includes smarter thinking devices and possibly their externalizations through which people who hold the principle, can implement its end.

The broad conclusions are as follows. AU might well be self-effacing when the ways of holding and applying the principle are restricted in a certain manner (say, to deduction), but AU itself does not involve such constraints. But because our world might have so constrained us, it might have been partially or wholly self-effacing. However, it has increasingly ceased to be so, as science and education have progressed and spread.[10] As the verdicts of utilitarianism depend on inputs from the real world, perhaps this world-dependent result on its self-effacement should be expected. However, this type of historical consideration is rarely found in the literature. The problem of self-effacement awaits more detailed scientific investigations, specifically into the trajectory of actual human history.

References

Akiba, T. (2016) What are the Problems (if any) of Self-Effacing Moral Theories?, *Applied Ethics*, 9, 12–29 [Written in Japanese].

Bowles, S. (2016) *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens* (Yale University Press).

Cox, D. (2012) Judgment, Deliberation, and the Self-effacement of Moral Theory, *Journal of Value Inquiry*, 46, 289–302.

Descartes, R. (1641) *Meditationes de prima philosophia*.

Gigerenzer, G. (2015) On the Supposed Evidence for Libertarian Paternalism, *Review of Philosophy and Psychology*, 6, 361–383.

Gigerenzer, G. (2007) *Gut Feelings: the Intelligence of the Unconscious*. (Viking).

Gigerenzer, G., Mata J., and Frank R. (2009) Public Knowledge of Benefits of Breast and Prostate Cancer Screening in Europe, *Journal of the National Cancer Institute*, 101, 1216–1220.

[10] Future researches need to investigate the role of freedom and democracy here.

Hansson, S. O., Risk, in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), URL = <http://plato.stanford.edu/archives/spr2014/entries/risk/>.

Japan Organ Transplant Network (2015) *News Letter*, 19, URL=<https://www.jotnw.or.jp/datafile/newsletter/index.html>.

Keller, S. (2007) Virtue Ethics is Self-Effacing, *Australasian Journal of Philosophy*, 85, 221–231.

Lessig, L. (1999) *Code and Other Laws of Cyberspace*. (Basic Books).

Mill, J. S. (1861) *Utilitarianism*.

Mill, J. S. (1859) *On Liberty*.

Parfit, D. (1984) *Reasons and Persons*. (Oxford University Press).

Rawls, J. (1971) *A Theory of Justice*. (Harvard University Press).

Sidgwick, H. (1907) *The Methods of Ethics*, 7th edition. (Macmillan).

Smith, M. (2001) Immodest Consequentialism and Character, *Utilitas*, 13, 173–194.

Stevens-Simon, C., et al. (1997) The Effect of Monetary Incentives and Peer Support Groups on Repeat Adolescent Pregnancies: a Randomized Trial of the Dollar-a-Day Program. *JAMA*, 277, 977–982.

Thaler, R., and Sunstein, C. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. (Penguin).

Williams, B. (1985) *Ethics and the Limits of Philosophy*. (Harvard University Press).