Blameworthy Bots

Don Berkich
Department of Humanities
Texas A&M University - Corpus Christi
6300 Ocean Drive
Corpus Christi, TX 78412
berkich@gmail.com

Abstract

An emerging consensus holds that fully morally blameworthy robots require a trio of cognitive capacities–awareness, understanding, and freedom of will–which presently exceed our engineering grasp by no small measure. Consequently we shelve puzzles over robots as full moral agents in favor of the problem of designing moral normative constraints on machine behavior. In this talk I argue that blameworthiness obtains even in the absence of awareness or freedom of will by investigating the cognitive grounds of *akrasia* or weakness of will so as to identify the minimal cognitive capacities robots must enjoy to be blameworthy. I conclude that blameworthy bots are much nearer on the horizon, from an engineering standpoint, than consensus allows.

Introduction

In an earlier talk[1] I argued that a shadow of the problem of original intentionality, which I dubbed *the problem of original agency*, threatens the prospects for genuinely autonomous robots. The argument is straightforward:

1. Necessarily, if X is an agent then X has original agency.
2. Necessarily, if X is designed and programmed, then X has only derived agency.
3. Necessarily, if X has only derived agency then X does not have original agency.
4. Every machine must be designed and programmed.
5. Therefore, no machine can be an agent.

The gist of the argument is that machines at most recapitulate the agency of their designers and users insofar as machines have no *self*-control whatsoever. The upshot for my present project is that any blameworthiness accrues not to the robots that at most transparently serve as causal mechanisms but to the designers and users who create and employ them as mere extensions of their own agency: The surveillance robot set to record and relay images upon detecting proximal motion is not responsible for violating privacy, those deploying it are. We must ask, then,

At what point in increasing sophistication do robots cease to be mere causal conduits and become themselves blameworthy?

Note that although Wallach and Allen give my puzzle a nod in their closing chapter,[14] it goes beyond their project of providing moral normative constraints on robot behavior. In part

[1] ”The Problem of Original Agency: On the (Im)possibility of WALL-E”, given at NA-CAP@IU 2008: The Limits of Computation

this is because the presumed requisite capacities so far exceed our current engineering expertise that full-bore blameworthiness seems a hopelessly premature topic. Moor, for example, gestures at consciousness, understanding, and freedom of will as the necessary cognitive capacities of what he calls "full ethical agents".[10, p. 20] Yet we haven't the faintest idea how to understand phenomenal consciousness within a physicalist framework in the first place, never mind having computational models of it. Much the same can be said for understanding and freedom of will.

I submit that another reason for drawing the investigation of robots up short of full ethical agents is the concept of blameworthiness itself. That is, it is unclear precisely *which* cognitive capacities blameworthiness presupposes. My purpose here is to try to make some progress on that question.

I start by assuming that there are fully blameworthy agents and we adult human agents, employing our ordinary cognitive resources, are just such agents. Under the further assumption of (token) physicalism, it follows that evolution has provided an existence proof that the cognitive prerequisites of blameworthiness are physically realizable. Finally, I assume that I am an agent minimally responsible for my actions at least insofar as I have the capacity for *self-control.* While I grant that none of these assumptions is uncontroversial, they are critical to my having any purchase on the question at hand. In particular, the assumption that responsible agency presupposes a capacity for self-control suggests a useful methodological strategy: Lift a page from Austin's "A Plea for Excuses" and investigate not self-control *per se* or *enkrateia* but failure of self-control, or *akrasia.*[9, p. 4] Thus,

> [T]o examine excuses is to examine cases where there has been some abnormality or failure: and as so often, the abnormal will throw light on the normal, will help us to penetrate the blinding veil of ease and obviousnesss that hides the mechanisms of the natural successful act. It rapidly becomes plain that the breakdowns signalized by the various excuses are of radically different kinds, affecting different parts or stages of the machinery, which the excuses consequently pick out and sort out for us. Further, it emerges that not *every* slip-up occurs in connexion with *every*thing that could be called an 'action', that not every excuse is apt with every verb–far indeed from it: and this provides us with one means of introducing some classification into the vast miscellany of 'actions'. If we classify them according to the particular selections of breakdowns to which each is liable, this should assign them their places in some family group or groups of actions, or in some model of the machinery of acting.[1, pp. 179-180]

To be sure, my goal is not to resurrect natural language philosophy. Rather, much as Austin seeks to understand agency by investigating when it goes awry, I seek to understand the mechanisms of self-control by investigating how we, seemingly paradoxically, can deliberately not exercise it. To wit,

> What in Anglo-Saxon philosophical circles is called the problem of weakness of will concerns what worried Socrates: the problem of how an agent can choose to take what they believe to be the worse course, overcome by passion. The English expression would not, or at least not primarily, bring this sort of case to mind, but rather such examples as dilatoriness, procrastination, lack of moral courage and failure to push plans through. The Greek word 'akrasia', on the other hand, means 'lack of control', and that certainly suggests the Socratic sort of example. [8, p. 97]

Largely by reference to the philosophical development of the problem of akrasia as it applies to human agency, I proceed by examining the range of what I shall call *akratic breaks* so as to identify the cognitive capacities they require. I conclude by evaluating Moor's requirements of consciousness, understanding, and freedom of will in light of the capacities my analysis uncovers. First, though, a response to Strawson is in order.

Participant Reactive Attitudes

Intuitively, *blaming* presupposes *blameworthiness.* In the absence of excusing conditions–unavoidable black-ice on the road, say–we blame the driver for the accident by reason of the driver's own recklessness. Yet if some version of psychological or causal determinism is true, the driver could not have but been reckless, in which case incompatibilism between determinism and the social practice of blaming threatens. Incompatibilism notwithstanding, we might call this 'intrinsic' justification for blame.

An alternative, compatibilist approach is to justify blaming not in terms of the relevant cognitive features of the agent, but instrumentally. Even if the driver could not have but been reckless, blaming him appropriately may mitigate future accidents by restricting, say, whether he can drive at all. Thus we blame because we find it useful to do so, not because we find the agent culpable. Call this 'extrinsic' justification for blame.

Strawson[13] argues that both intrinsic and extrinsic justification for blame miss the point. Blaming is one of many participant reactive attitudes which require no justification beyond their expression of our expectations as members of a social community. Blaming does not presuppose blameworthiness; rather, blameworthiness presupposes the social practice of blaming, set against the backdrop of being a member of a social community rich with participant reactive attitudes. Of course, there are exceptions. Children and adults with severe mental disorders merit objective attitudes, but only because they are not yet or not currently full participants in the social community.

To be sure, I have no particular quarrel with Strawson's approach, provided we recognize its limited application. For while Strawson arguably manages to sidestep the problem of whether it makes any sense to hold a determined agent responsible, it comes at the cost of having any way of dealing with robots or, say, aliens. Since blameworthiness presupposes the social practice of blaming in a community, robots and aliens, failing to be part of the community of any human persons in the first place, can never be blameworthy. Whatever its advantages, Strawson's approach hobbles any question we might ask about whether a given robot or alien is blameworthy. My question in particular about the point at which we will need to recognize the blameworthiness of the robots themselves becomes otiose under Strawson's analysis. If they were already full participants in our community, then we as a result would take them to be blameworthy. Presumably we welcome other adults to be full participants because they are sufficiently like us, making not doing so capricious, at best–not so for robots or aliens.

Were aliens to visit, however, I think it would make sense to wonder about whether and to what extent we are justified in holding them responsible, antecedent to, and perhaps as a prerequisite on, their being welcomed into our social community. Similarly, I submit that the question of blaming robots hinges on their being blameworthy and not merely causal conduits for others' agency. In short, we require intrinsic justification for blaming robots, if such is possible.

The Early Puzzle

To recap: Blaming, then, presupposes blameworthiness, or at least we must assume as much where robots are concerned. Further, I've assumed that a robot is blameworthy insofar as it is responsible for its actions in at least the minimal sense in which it is not a mere causal conduit. That is, a blameworthy robot necessarily has the capacity for original agency. Original agency, though, presupposes self-control. Our question then becomes, at what point will a robot have the cognitive sophistication necessary to exercise self-control? My strategy, following Austin, is to uncover the cognitive prerequisites of self-control by considering when normally self-controlled agents, ourselves, lack it, which brings us to the problem of akrasia.

The nicotine addict huddled against the freezing Massachusetts winter to grab a smoke, although often taken to be the paradigm of the weak-willed, poses no special problem of akrasia. She has, after all, a strong overriding reason for braving the elements: Her addiction. Even if extremely conflicted, she nonetheless enjoys enkrateia.

No, the puzzle of akrasia arises when we imagine an agent who intentionally pursues a course of action which diverges from their own considered judgment. The politician who violates his wedding vows to have an affair with a staffer, the student who plagiarizes portions of a term paper, the research scientist who fudges data, and the professor who watches television instead of getting caught-up on grading all present a similar challenge: How is it possible that they intentionally acted in ways they themselves judge worse than an alternative they believed open to them? The apparent absurdity compelled Plato to reject the very notion that reason could come apart from itself in this way, since [n]o one, who either knows or believes that there is another possible course of action, better than the one he is following, will ever continue on his present course[12, *Protagoras* 358b-c].

Aristotle derides Plato's conclusion, since it contradicts the plain phenomena.[4, *Nicomachean Ethics* 1145b27] The challenge is to explain how an agent can intentionally pursue an akratic alternative, contrary to his own better judgment. Aristotle's approach is to point out that the agent's practical judgment is grounded in competing and incompatible practical syllogisms.[2] That is, the agent can have reasons without using them, since in the case of competing practical syllogisms, only one can issue in action.

> ...human beings may have knowledge in a way different from those we have described. For we see that having without using includes different types of having; hence some people, such as those asleep or mad or drunk, both have knowledge in a way and do not have it. Moreover, this is the condition of those affected by strong feelings. For spirited reactions, sexual appetites, and some conditions of this sort clearly [both disturb knowledge and] disturb the body as well, and even produce fits of madness in some people. Clearly, then [since incontinents are also affected by strong feelings], we should say that they have knowledge in a way similar to these people. [4, *Nicomachean Ethics* 1147a10-18]

The strong feeling submerges or occludes the competing practical syllogism in favor of the non-akratic alternative in such a way that the akrates can be said to have the knowledge without being compelled to act on it. Instead, the akratic alternative is more present to the akrates inasmuch as the strong feeling brings to the foreground and makes vivid its justifying practical syllogism.

> For one belief is universal; the other is about particulars, and because they are particulars, perception controls them. And in the cases where these two beliefs

[2] See [2] for a more comprehensive discussion of Aristotle's solution.

> result in one belief, it is necessary, in one case, for the soul to affirm what has been concluded, but, in the case of beliefs about production, to act at once on what has been concluded. If, for instance, everything sweet must be tasted, and this, some one particular thing, is sweet, it is necessary for someone who is able and unhindered also to act on this at the same time.
>
> Suppose, then, that someone has the universal belief hindering him from tasting; he has the second belief, that everything sweet is pleasant and this is sweet, and this belief is active; but it turns out that appetite is present in him. The belief, then, [that is formed from the previous two beliefs] tells him to avoid this, but appetite leads him on, since it is capable of moving each of the [bodily] parts.
>
> The result, then, is that in a way reason and belief make him act incontinently. The [second] belief is contrary to the correct reason, but only coincidentally, not in its own right. For the appetite, not the belief, is contrary [in its own right to the correct reason]. [4, *Nicomachean Ethics* 1147a24-1147b3]

The strong appetite for the sweet overwhelms the correct reason the akrates *has* but does not *use*. Akrasia is possible, then, because the subversive strength of the desires associated with the reasons behind pursuing the akratic alternative temporarily submerge the akrates' better reasons for doing otherwise.

The Late Puzzle

Austin famously points out, however, that there need be no such strong desire for the akratic alternative:

> I am very partial to ice cream, and a bombe is served divided into segments corresponding one to one with persons at High Table: I am tempted to help myself to two segments and do, thus succumbing to temptation and even conceivably (but why necessarily?) going against my principles. But do I lose control of myself? Do I raven, do I snatch the morsels from the dish and wolf them down, impervious to the consternation of my colleagues? Not a bit of it. We often succumb to temptation with calm and even with finesse.[1, p. 198]

Indeed, akrasia seems less problematic when compelled by strong desire, as Aristotle would have it, yet simultaneously more common and more problematic when not so compelled. Casting it as a logical problem, Davidson suggested that we have three individually plausible but jointly incompatible principles:

> P1. If an agent wants to do x more than he wants to do y and he believes himself free to do either x or y, then he will intentionally do x if he does either x or y intentionally.
>
> P2. If an agent judges that it would be better to do x than to do y, then he wants to do x more than he wants to do y.
>
> P3. There are (akratic) actions.[5, p. 23]

Where,

> In doing y an agent acts (akratically) iff
>
> i. the agent does y intentionally
>
> ii. the agent believes there is an alternative action x open to him
>
> iii. the agent judges that, all things considered, it would be better to do x than to do y.[5, p. 22]

That the akrates does y intentionally seems incompatible with P1 and P2 together, given his belief that there is a better alternative x open to him. The curious phrase, "all things considered", blocks the looming contradiction since P2 requires an all-out judgment–that is, *sans phrase*–that it would be better to do x than to do y. Akrasia is possible insofar as the akrates has failed to make an all-out judgment in such a way that judgment would be closely coupled to subsequent intentional action. Thus in the akrates, judgment decouples from subsequent intentional action since the akrates also makes an all-out judgment on behalf of the akratic alternative.[3] Thus,

> ...the way could be cleared for explanation if we were to suppose two semi-autonomous departments of the mind, one that finds a certain course of action to be, all things considered, best, and another that prompts another course of action. On each side, the side of sober judgement and the side of incontinent intent and action, there is a supporting structure of reasons, of interlocking beliefs, expectations, assumptions, attitudes, and desires.[6, p. 181]

To be sure, Davidson's approach rules out the possibility of all-out, *sans phrase* akrasia. In cases of last-ditch akrasia [11] or strict akrasia [9], the akrates has made an all-out judgment that it would be better to do x than y, and thus cannot be construed as being of two minds on the matter. It seems must in that case reject P2, since an all-out judgment that one alternative is better than another must not necessarily motivate in the way we think it ought.

What Akrasia Shows

My discussion of akrasia in no way does justice to the richness of the philosophical discussions or the many approaches that have been taken to it. Rather, I have sought to describe a few of the highlights with hopes of being able to draw some conclusions about the cognitive capacities akrasia presupposes go awry.

Akrasia is first and foremost a disconnect between the agent's intentional action and the reasons she has for acting, where her reasons for a non-akratic alternative fail to trump–or, indeed, have any effect on whatsoever–her intentional pursuit of the akratic alternative. Different solutions describe how the akratic break between reasons and intentional action occurs. Aristotle's account is attractive, inasmuch as it locates the akratic break in reason itself. The better reasons the agent has for intentionally acting in one way are obscured by the strength of feeling associated with the reasons the agent has for the lesser alternative. Davidson's account distinguishes between the kinds of judgments an akrates makes, which are conditional on the reasons the agent happens to have at the time, and the illicit all-out judgment the agent draws which leads to intentionally pursuing the akratic alternative. Strict akrasia of course precludes this sort of explanation.

[3] For a more comprehensive and critical discussion of Davidson's approach, see [3].

If we blame someone for intentionally doing what they themselves admit they should not have done, then blameworthiness is underwritten by the same cognitive capacities as akrasia. Breaking it down, the agent must be capable of acting intentionally and, thus, of forming intentions to so act. Further, the agent must be capable of evaluative deliberations so as to ascertain what is the best course of action. In the self-controlled agent, the agent deliberates about the best course of action, makes a judgment about the best course of action, forms an intention to so act as a result, and thereby intentionally acts. In the self-controlled agent, that is, reasons generate actions as they ought. Notice we might still blame the self-controlled agent. They could simply have made a mistake as to the best course of action. Nevertheless, making a mistake in deliberation versus correctly deliberating yet not following through warrant different degrees of blame.

Yet is the akrates blameworthy *simply* by failing to exercise self-control, or are freedom of will, understanding, and phenomenal consciousness also required, whether to have failed to exercise self-control in just the right way, or otherwise?

Complicating matters considerably is the fact that it is not at all clear what Moor or Wallach and Allen have in mind when they gesture at freedom of will, understanding, and phenomenal consciousness as prerequisites to full moral agency. There are, for example, a number of ways to cast freedom of will, some of them compatible with determinism, some not. Thus the arguments I present below decoupling free will, understanding, and phenomenal consciousness from blameworthiness are at most suggestive, since alternate accounts are plentiful. So as to accommodate as many such accounts as possible, I shall assume as little as possible.

Consider each in turn.

Freedom of Will

Note that on Davidson's (and most others') account of akrasia, it suffices that the agent merely believe the non-akratic alternative is open to him. It is not that the agent knows the non-akratic alternative is available or that the non-akratic alternative simply be available. Thus it need not be the case that the agent be free to do otherwise, consonant with Frankfurt's compatibilist arguments.[7] Alternate courses of action need not be open to the blameworthy agent in the metaphysically demanding sense that he could have done otherwise. Instead the blameworthy agent need merely meet the weaker epistemic condition of having believed that he could have done otherwise. One can imagine, for example, a willing yet unwitting kleptomaniac who steals because she wants to steal, but who, were she to have chosen otherwise, could not have but stolen given her as yet undiscovered compulsion. Stealing because she wanted to makes her blameworthy, regardless of her lack of alternatives.

Understanding

Understanding is more difficult. If what we mean by 'understanding' as a requirement of blameworthiness is that the agent have consciously chosen the blameworthy course of action, then it is unclear how understanding as a requirement is anything but redundant on phenomenal consciousness. If, on the other hand, 'understanding' is a distinct requirement, then at most it is the requirement that the agent's deliberative bases bear the mental content we expect them to in order to play the role they do in the mental economy underwriting agency. The question of what it is also like for the agent to be in those mental states is another, more difficult, requirement. For certainly the blameworthy agent, insofar as he has self-control, has a sufficiently robust mental economy to support practical deliberation.

Phenomenal Consciousness

The sticking point, then, would appear to be phenomenal consciousness. That is, does blameworthiness presuppose conscious awareness since, in the words of an anonymous referee, "[b]laming a machine or carbon zombie is like blaming a rock that you stub your toe on"? Maybe so, but would those intuitions survive a precocious rock that relentlessly positions itself to optimize toe stubbing opportunities?

To be sure, it would not do to point out that we sometimes apportion posthumous blame, since the blame is for prior actions taken which may well have depended on consciousness. There are, however, cases where it seems appropriate to blame despite the absence of conscious awareness. The possibility of unconscious beliefs and desires in human agents, along with the possibility of acting on them intentionally without recognizing their deliberative grounding (perhaps by acquiring a *post hoc* justification), suggests that the blameworthy agent need not be aware of their own cognitive states as they play their roles in deliberation.

To be sure, we may initially balk at the notion of blaming a philosophical zombie. After all, what good could come of it? We blame because we think that the target of our blame will reflect on their bad acts and feel some remorse. Nevertheless, an agent surely can be blameworthy without any subsequent blaming ever accomplishing anything. Further, our investigation of akrasia was framed entirely in terms of the role of various mental states in the akrates' mental economy. Nowhere did we need to draw on what it was like for the akrates to be in those mental states in examining failure of self-control.

Concluding Remarks

The cognitive capacities presupposed by blameworthiness and which justify the practice of blaming turn out on reflection to be relatively modest: Evaluative deliberation, intention formation, and intentional action jointly suffice. Granted, these capacities are still a tall order for computational modeling.

Where Moor assumes we must have phenomenal consciousness, understanding, and freedom of will to have responsible robot agents, on analysis we require at most the cognitive grounds of agency. It remains a difficult question whether those grounds are computationally tractable, but I take it that there is greater cause for optimism without having to worry about freedom of will or phenomenal consciousness. Contrary to consensus, then, we may find ourselves justifiably blaming bots surprisingly soon.

References

[1] J.L. Austin. *A Plea for Excuses*, pages 175–204. Philosophical Papers. Oxford University Press, Oxford, 3rd edition, 1979.

[2] D. Berkich. On two solutions to akrasia. *Philosophical Writings*, 33:34–52, 2006.

[3] D. Berkich. A puzzle about akrasia. *Teorema*, 26(3):59–71, 2007.

[4] Translated by Terence Irwin. *Aristotle: Nicomachean Ethics.* Hackett Publishing Co., Indianapolis, 2nd edition, 1999.

[5] D. Davidson. *How is Weakness of the Will Possible?*, pages 21–42. Essays on Actions and Events. Clarendon Press, Oxford, 1980.

[6] D. Davidson. *Paradoxes of Irrationality*, pages 289–305. Philosophical Essays on Freud. Cambridge University Press, Cambridge, 1982.

[7] H.G. Frankfurt. Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23):82939, 1969.

[8] J. Gosling. *Weakness of the Will.* Routledge, London, 1990.

[9] A.R. Mele. *Irrationality: An essay on akrasia, self-deception, and self-control.* Oxford University press, Inc., Oxford, 1987.

[10] J. H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):1821, 2006.

[11] D. Pears. Motivated irrationality. *Proceedings of the Aristotelian Society*, 56:156–78, 1982.

[12] Plato. *Protagoras*, pages 308–352. The Collected Dialogues of Plato. Princeton University Press, Princeton, 1961.

[13] P.F. Strawson. Freedom and resentment. *Proceedings of the British Academy*, 48:1–25, 1962.

[14] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong.* Oxford University press, Inc., Oxford, 2010.