Responsibility, Applied Ethics, and Complex Autonomy Theories

When I was kindly invited to contribute a paper to this volume, the letter of invitation included the following sentence:

As you know, the twin concepts of autonomy and identification have become increasingly important within contemporary philosophy, especially in discussions of moral responsibility and applied ethics.

It seems to me that the claim that the concepts of autonomy and identification – or perhaps the concept of autonomy which is often related to identification- are important for moral responsibility and applied ethics is often made too quickly or without some needed reservations. While some things worthy of the name 'autonomy' are clearly important in moral responsibility and applied ethics, the 'twin concepts' of autonomy and identification as studied by many philosophers in the Frankfurt-inspired tradition in moral psychology do not have *clear,* transparent relevance for moral responsibility or applied ethics. That is, although some broadly-Frakfurtian theories of autonomy and identification may very well have major implications for moral responsibility and applies ethics, one cannot – perhaps can no longer - assume uncritically that any theory of autonomy, simply by virtue of being a theory of autonomy, has significant implications for moral responsibility or applied ethics, and even in cases in which it is plausible to think there would be significant implications, it is often not clear exactly what these

implications are. In arguing for this conclusion, I will first focus on moral responsibility first, and then say something about applied ethics.

Let me explain what I mean by "clear, transparent relevance". Some views of autonomy, such as Fisher and Ravizza's view, are driven by the assumption the class of deeds for which we are morally responsible and the class of deeds that are instances of autonomous agency are identical, or approximate each other quite closely. Often, they treat 'autonomous action' and 'action for which we are morally responsible' as interchangeable. A philosopher advancing such a view will be guided by intuitions about moral responsibility in a fairly clear way: she will assume by default that if an agent would commonly be held responsible for an action, or praiseworthy or blameworthy for it, the action is, prima facia, an instance of that agent exercising autonomy, while if a person would commonly be exempt from moral responsibility, it would be prima facia counter-intuitive to regard him as autonomous. In any case of an apparent clash between the author's definition of autonomy and common intuitions about moral responsibility, it will be her natural task to provide a satisfactory explanation of this apparent clash. One would expect such a philosopher to make some reference to all paradigmatic cases of humans who are not morally responsible for something that one is usually responsible for – psychotics, victims of coercion, etc. Within the framework of such a view, the claim that an action is or is not autonomous has a fairly obvious relevance to issues of moral responsibility: an autonomous action, for all or almost all intents and purpose, simply *is* an action for which we are morally responsible.

But things are more complicated with theories of autonomy or identification whose authors neither argue nor assume that the class of autonomous actions is identical (or

near identical) to the class of actions for which we are morally responsible. These theories are the subject of this paper. There are many theories who do not make that assumption, starting perhaps with Frankfurt's introduction of the idea of identification: early Frankfurt thinks of identification –and, we assume, of autonomy - in terms of acting on desires that you desire to act on, but has no intention of arguing, strictly speaking, that I am exempt from moral responsibility if I desire to act on my desire to prepare my class but act instead on my 'outlaw desire' to read *The Curious Incident of the Dog in the Nighttime*. Early Frankfurt holds that there is a significant connection between his notion of autonomy and responsibility, but unlike Fisher and Ravizza, he does not assume that the set of actions for which we are responsible and the set of actions that are autonomous are identical or near-identical. So divorced is Frankfurt from this assumption that he sees no need to address at any length what may seem, to the old-fashioned free will theorist, like an apparent conflict between intuitions about moral responsibility ("I am morally blameworthy for not preparing for my class") and implications of his view ("I am not autonomous if I akratically fail to prepare for my class"). Later Frankfurt, who thinks of autonomous actions as wholehearted actions, clearly does not intend to argue that we are only responsible for wholehearted actions. His work is driven not by intuitions about moral responsibility, but about internal conflict and the disturbing sense that we sometimes get of being acted-upon by our own minds.

To a greater or lesser degree, many theories of autonomy and identification seem to imply that the set of autonomous actions is significantly smaller than the set of actions for which we are commonly held morally responsible. While their authors are often inclined to hold that some connection exists between autonomy and moral responsibility, they do not simply assume that their account of autonomy must serve directly as an account of the

necessary conditions for moral responsibility, and if an action appears non-autonomous by their lights but actionable by ordinary standards they do not see it as a philosophical emergency requiring immediate action by way of in-depth explanation. Consider, for example, a claim advanced by Ekstrom in this volume:

"What I am suggesting, as one inviting way to view autonomous action, is that a person acts autonomously only when acting from motivation of a certain sort: one that (i) has undergone critical evaluation with respect to his conception of the good, (ii) was uncoercively formed, and (iii) coheres with his other acceptance and preference states. It may be that we regularly live with internal conflict, but that we act autonomously only when acting on one of those attitudes that is central to who we are, where centrality to the self is established as I have described."

These are fairly strict conditions for autonomy. If we "regularly live with internal conflict", it very well may be that autonomy, as Extrom defines it, is a bit like happiness or fitness: an eminently reasonable things to which to aspire, but not the default condition of human beings. It may be even more rare, depending on the strictness of our interpretation of "critical evaluation with respect to his conception of the good" – something that even people whom we would not regard as particularly conflicted don't do very often. As Ekstrom herself admits in a footnote, her account of autonomy, if coupled with the assumption that autonomous actions are actions for which we are morally responsible, would yield an account of moral responsibility that is far too restrictive (it will probably exclude, for example, the main character of "Crime and Punishment", with his torn Russian soul, as responsible for the murder he commits). But this is not a problem for Ekstrom's paper, because she does not argue that we are only

responsible for actions that meet her conditions. The question of the precise relevance of the quoted view to moral responsibility is left quite open.

Let me refer to theories of autonomy that cannot easily dispense with the expression "autonomous action" and replace it with "action for which we are morally responsible" as *Complex Autonomy Theories*. I do not wish to argue that CATs, or any particular CAT, are not relevant to moral responsibility, but that the clichés 'autonomy grounds moral responsibility' and 'autonomy is central to moral responsibility' are not automatically true for CATs. It cannot be simply taken for granted that a CAT is relevant to moral responsibility, much less that it "grounds" moral responsibility: anyone who wishes to argue that a CAT is relevant to moral responsibility can be fairly asked to defend and clarify her view. Perhaps in some cases the defense and clarification would be much easier than in others, and I am aware of the fact that "complexity" or "distance from the assumption that autonomous action is an action for which we are morally responsible" are matters of degree. But if a theory of autonomy is a CAT, defending and clarifying the connection between is and moral responsibility will not be as easy as saying "as we all know, autonomy theories has implications for moral responsibility": it will be a subject for further theorizing, and at times it may even amount to a research program all on its on. This is true however valuable the CAT may be in other, non-responsibility-related ways. After all, a CAT can be valuable and interesting even if it has nothing to do with moral responsibility or, for that matter, with applied ethics. There is more to the philosophical study of the vagaries of the human heart than what is contained in these subjects.

Two immediate objections to my claim about CATs and moral responsibility comes to

mind. First, an objector could claim that while she is not simply describing the class of actions for which we are moral responsibility, her CAT clearly describes the class of actions for which we are *directly* morally responsible, and therefore is linked to moral responsibility in a clear-enough way. For example, if a theory implies that procrastinating against one's best judgment is not autonomous, and is confronted with the intuition that such procrastinating can still be blameworthy, the theorist may explain that one is not directly responsible for the procrastinating, but one is blameworthy for autonomously taking a course of action that allowed the actions to happen. Secondly, an objector may claim that if a theory of autonomy successfully describes a properly that is essential to being a person and is not shared with non-persons, it is thereby relevant to moral responsibility, which only persons have. In other words, instead of assuming an identity between the class of autonomous actions and the class of actions for which one is responsible, one can appeal to an identity between the class of autonomous *agents* and the class of agents that can be held morally responsible, and this establishes a clear enough connection between autonomy and moral responsibility. I will address these objections in order.

The "Indirect" Strategy

It is possible for a philosopher to claim that her CAT clearly describes the class of actions for which we are *directly* morally responsible, and therefore is linked to moral responsibility in an clear-enough way. I do not doubt that such a strategy could work. That does not mean, however, that saying, offhandedly, something like "of course you are responsible for procrastinating, but it's indirect responsibility – you should have controlled yourself more strictly" is enough to establish a clear connection between a CAT

and moral responsibility. It may take serious argument to show that such an "indirectness" claim is true of all apparently non-autonomous actions for which we are apparently morally responsible. Beyond such intuitions as "you should have controlled your temper", it can also be difficult to clarify what exactly the meaning of such an indirect responsibility claim is, and it needs to be fleshed out theoretically rather then taken to be a pre-theoretical assumption.

In "The Possibility of Practical Reason", Velleman gives us the following example of failure of autonomy:

I have a long-anticipated meeting with an old friend for the purpose of resolving some minor difference; but . . . as we talk, his offhand comments provoke me to raise my voice in progressively sharper replies, until we part in anger. Later reflection leads me to realize that accumulated grievances had crystallized in my mind, during the weeks before our meeting, into a resolution to sever our friendship over the matter at hand, and that this resolution is what gave the hurtful edge to my remarks. In short, I may conclude that desires of mine caused the decision, which in turn caused the corresponding behavior; and I may acknowledge that these mental states were thereby exerting their normal motivational force, unabated by any strange perturbation or compulsion. But do I necessarily think that I made the decision or that I executed it? Surely, I can believe that the decision, though genuinely motivated by my desires, was thereby induced in me but not formed by me; and I can believe that it was genuinely executed in my behavior but executed, again, without my help. Indeed, viewing the decision as directly motivated by my desires, and my behavior as directly governed by the decision, is precisely what leads to the thought that as my words became more shrill, it was my resentment speaking, not

I. (Velleman 1992, 464-65)

Velleman takes my yelling at my friend not to be a full-fledged action, or an autonomous action. Common sense holds that I deserve blame for my rudeness, and for many other deeds that would be, on similar grounds, regarded as non-autonomous by Velleman. Thus, Velleman's theory seems to postulate that the class of autonomous actions is considerably smaller than the class of actions for which we are morally responsible, and thus it seems that I would be committed to saying that therefore, the theory's implications for moral responsibility are far from obvious. In a footnote, Velleman says that he does not mean to deny that I am morally responsible for the incident, but he suggests that there is still an obvious connection between his view of autonomy and moral responsibility by claiming that I have a duty to be vigilant about "unconsidered intentions" and actively prevent them from running loose, and so I am blameworthy not for yelling, but for failing to prevent myself from getting into the state in which I find myself. I, the agent, have autonomously failed to keep good watch on my resentment.

Let me take Velleman's footnote as an example of the way in which an indirect responsibility claim can be used to establish an obvious connection between a CAT and moral responsibility, and use it to demonstrate why I think that the claim that all non-autonomous actions for which we seem to be responsible are actions for which we are indirectly responsible is not self-evident.

My first problem with the Velleman's indirectness claim is one that I have handled in details elsewhere, and I will state it briefly. It involves the complexities of cases in which we seem praiseworthy, rather than blameworthy, for the non-autonomous action.

Suppose we change Velleman's case and shift from talk of blame to talk of praise. Imagine that the "accumulated grievances" that crystallized in my mind into a decision to break off my friendship involve my friend's increasingly immoral behavior, (and the increasing moral dubiousness of being allied with him) which I have been consciously ignoring or downplaying or underestimating. There are occasions on which such a breakup marks a pivotal moral step for a person, or at least an occasion to say "good for you", and hence warrants moral praise. It may still be occasions for moral praise even though, as I walk away from the meeting, I tell myself that I should have done it long ago, and that ideally it would have been better if I have done it in a more dignified and pre-planned manner. Where does this praise come from? No story about a duty of vigilance can easily explain it. I deal with praiseworthiness cases in detail in *Unprincipled Virtue.*

Furthermore, even in cases of blame, the 'vigilance' thesis needs clarification and defense. There are many cases in which we appear to be blameworthy for actions that are a lot like the one described by Velleman – surprising to the person who performs them and involving akrasia and a sense of alienation. People who succumb to rage, temptation, or visceral inhibition - adulterers, impulsive aggressors, akratic procrastinators, and so on- often say things that suggest alienation, "I don't know what got into me" and "the devil made me do it". Though we usually "know what they mean", just as we know what Velleman means when he says "it was my resentment speaking not I", we usually hold them blameworthy. Are all of these cases of indirect responsibility? What, in such cases, is the course of action for which the sinner is *directly* responsible?

To be sure, there are some cases in which "being vigilant" is the answer. There are many occasions on which we do hold people blameworthy for failing, as it were, to check their

mental brakes. If, for example, I know that drinking a large espresso or missing a dose of lithium is likely to make me irritable, and I have yelled at my friend as a result of neglecting to watch my coffee or lithium intake, it is quite plausible to say that my guilt consists in my negligence. Even if no such stark mechanism is in operation, it may be that, as I felt my blood pressure go up, I should have taken a deep breath and counted to ten. But things become less clear at this point. It is quite possible to imagine a scenario in which no such "count to ten" measures were available to me, or no effective ones anyway. It is also sometimes the case that the agent could not be expected to know of such measures in time to use them (perhaps powerful aggressive urges have never appeared in me before, and when such an urge appears it takes me by such surprise that I do not notice it until I am already screaming). There are also many cases in which the agent has already taken such measures, and in general tried as hard as she could not to follow her "outlaw" desires, but her attempts and measures fail. In many such cases, we still blame the unautonomous aggressor (or the akratic adulterer, procrastinator, etc).

In lieu of the simple 'vigilance' thesis, one might suggest here that the autonomous course of action for which the agent is blameworthy must be some sort of failure that resulted in her current state of weakness of will – resulted in the fact that "as hard as she can" is not very hard. Thus, the agent must be blameworthy for having failed to perform some character-building action, or having knowingly performed some character-eroding actions. This view would have similar problems to the vigilance view when it comes to cases of non-autonomous behavior which is putatively praiseworthy (as in the cases of Huckleberry Finn and Oscar Schindler, which I discuss elsewhere). Other than that, the main problem I see with the character-development view has to do with the fact that it necessitates a picture of human life in which we have an incredible amount of control

over our characters – an amount of control that most parents only wish they could have over the development of their children's characters. How often do we knowingly and autonomously perform character-building or character-ruining actions? To be sure, occasionally we do. Mr. Tucker, a character in Christopher Buckley's satire *The White House Mess*, knows that entering the White House is likely to turn him into what he calls "a jerk." Yet, he chooses to enter the White House, and his moral character is in fact harmed in the ways in which he predicted it would be. From Balzac's Restiniaque to Trudeau's Michael Doonesbury, in fact, fictional characters can be found who make clear-eyed decisions in favor of courses of action which will gain them money or power but will harm their integrity or compassion. I do not doubt such decisions occur in real life. But instances of such decisions are rare – considerably rarer than the autonomy-oriented moral psychologist needs. Successful, intentional, character-building or character-ruining actions performed by a person upon himself are even more rare than successful New Year's resolutions. It is the exception, rather than the rule, that a person's character is substantially self-made, which is why a self-made good character is so impressive in the first place.[i] In many cases in which people lack self-control with respect to some of their desires (or when they simply do not have strong enough "good" desires to combat the "bad" ones), this weakness is primarily the result of early upbringing and all sorts of unintentional psychological reinforcement (and by the time one may think about changing one's character, it is already at least half-shaped). To the extent that agents contribute to the creation of their weaknesses by means of their autonomous actions, it is usually not in the straightforward way Tucker influenced his own character. Tucker knew about the way his White House job was likely to affect his character. Quite often, however, an agent chooses her character-shaping actions without any knowledge of the way in which they are likely to shape her character, and does so in circumstances in

which she could hardly be expected to know better (any parent trying to shape the character of a child knows how hard it is to make such predictions). One is not usually in a position to predict whether her choice of a job, school, marriage partner, friends, or area of domicile will affect her moral character in some fashion, not to mention the many choices which initially appear too insignificant to fuss about. Thus, it is quite unlikely that what unautonomous blameworthy agents are to blame for is *always and only* some autonomous failure of character-building.

To go back to the simple idea that a duty of vigilance or self control explains the blameworthiness of non-autonomous actions; there is a type of case in which there are independent reasons to believe that we are blameworthy by virtue of *having* a certain desire or motivational factor to the extent that we act on it at all, not by virtue of failing to control it. These are cases in which we act on sinister motives, where our reasons for action are in essential, rather than accidental, conflict with morality. Sexual desire, hunger, desire for money, and other traditional "temptations" are not, by themselves, sinister – they do not conflict with morality, though in the wrong set of circumstances, they can lead you to do something immoral. Thus if, for example, there are cases in which almost any human being, regardless of emotional makeup, would be moved by sexual desire – a motive that is morally neutral by itself – to the point of committing adultery, we may reasonably say that some adulterers, under such circumstances, and blameworthy not so much for their adulterous action but for leading themselves into those circumstances in the first place, or some similar failure of vigilance. Things are different, however, if our unautonomous sinner acts not from a neutral desire but out of a malevolent motive, such as sadism or racial hatred. If I lash at my friend because I relish the suffering of my fellow human beings, I am blameworthy even if I have done all in my

power to control and eradicate my sadism (compare "sorry, I haven't eaten for 24 hours and I couldn't help eating your chocolate" with "sorry, I haven't seen a person in pain for 24 hours so I couldn't help eating your chocolate"). Something similar seems to be true for 'slips' motivated by serious racial prejudice. As Hursthouse (1997) hints, the confession "I am utterly disgusted by Asian people but I am doing my best to control it" is a far cry better than a whole-hearted hatred of Asians, but is also a far cry from the confession of a morally perfect agent; the ego-dystonic racist who thinks that there is a lot to improve in her moral character, that she could be a better person, is after all, correct in her assessment: she *could* be a better person than she is (the same, I take it, holds for the "recovering" sadist).

One may object to this view by pointing out that heaping blame upon people such as the visceral racist or the visceral sadist makes the world a worse place, in that obsession with the blameworthiness of one's visceral feelings and desires tends to backfire.[ii] As long as a person knows that her visceral or unconscious desires are bad, the argument may proceed, tormenting her and sanctioning her for desires that she does not endorse may only lead to counterproductive results – no one, after all, can cope with being blamed or blaming herself all the time, and encouraging people to dwell too much on the badness of their visceral desires is likely to result in the activation of psychological defenses which are likely to interfere with them mending their ways. I mention this objection because I am at least partially in agreement with those who fear that attacking an involuntary sinner who already is trying to mend her ways is often a counterproductive – and therefore the wrong – thing to do. It may even be cruel, or at least unforgiving in a context where the virtuous agent would be forgiving. I would like to point out, however, that my view to the effect that the inadvertent sinner is blameworthy does not imply that

punishing her – even verbally – is a good thing for society to do or that obsessing over her sins is the right thing for her to do. As I argued elsewhere, to say that one is blame*worthy* is not to say that one *should be blamed*. In some circumstances, blame may be warranted without an expression of blame being morally desirable.

These are some consideration that make the claim that all responsibility for actions which are deemed non-autonomous by Velleman is indirect and derivative from responsibility to prior autonomous actions is a claim that needs clarification and defense, and thus if Velleman wanted to say that on his view, autonomy grounds moral responsibility he would have to argue for this conclusion and to make clear how it is supposed to ground it. It is easy to see how this same may be true for many other CATs.

CATS and other Animals

In away that is somewhat reminiscent for the Indirect Responsibility claim, one may argue that there is a clear-enough relationship between CATs and moral responsibility in that CATs capture things that tell apart persons from other creatures, and as only persons are morally responsible (and, some would say, all persons, though this is more complicated), then CATs tell us what it is to be a responsible agent – creature capable of morally accountable action. It is true that humans are the only morally responsible creatures we know at the moment, and that typically, CATs identify mental conditions that only a human is likely to have. CATs focus on forms of inner hierarchy and/or inner struggle which seem to exist in all 'normal', adult humans and in no other creature.

But this is not enough to make it a trivial assumption that CATs are relevant to moral

responsibility. It is not enough because 1) many mental abilities are uniquely human, and not all of them are clearly relevant to moral responsibility and 2) it seems prima facia possible to explain a lot of things about the non-responsibility of animals without appealing to any CAT.

Let me start from the second point. In *Leviathan*, Hobbes succinctly gives the following view:

> To make covenants with brute beasts is impossible, because not understanding our speech, they understand not, nor accept of any translation of right, nor can translate any right to another: and without mutual acceptation, there is no covenant.

Hobbes is only speaking here of contracts.[1] But if we wish to explain why animals are not moral subjects, it may be an interesting exercise to see how far such commonsense facts as animals not understanding our speech can take us, before we have reached anything quite as complicated as agent-autonomy. The exasperating fact that your cat cannot understand your request that she be careful in handling your computer keyboard from now on counts for a lot when you remind yourself that she is exempt from moral responsibility for knocking it off the table again. Now let us expand the Hobbesian notion of "not understanding our speech" and speak simply about things that animals, given their intellectual capacities, do not understand. Let us again consider a situation in which

[1] Though the ability to make contracts and being a moral subject are very similar things for him.

we are tempted to blame a non-human animal but think better of it. A child discovers that the family dog destroyed her dinosaur-shaped toy. She becomes angry; "But it's *my favorite dinosaur*!" she screams. We may well imagine a parent explaining to her that "He's only a dog, darling. He does not understand that it's your favorite dinosaur." The dog does not understand *mine*, *favorite* or *dinosaur*, not even in the murky, visceral way in which a small child does. Similarly, the dog's mind presumably cannot grasp – nor can it track, the way even unsophisticated people can – such things as increasing utility, respecting persons or even friendship. As Hobbes hints, even if some proto versions of these notions exist in the animal's mind, these are not concepts that it can sophisticatedly apply to humans. Thus, even if this animal can act for reasons, to some extent, it cannot respond to *moral reasons*, which makes it very hard to regard it as blameworthy. To judge a dog vicious for not responding to moral reasons would be similar to judging a dog a philistine for not being able to appreciate Mahler. Dwelling on this banal list of things that dogs cannot understand shows us the possibility that what prevents dogs from being moral subject may or may not have to do with things like having second order desires (or whatever your favorite notion of autonomy). The connection is far from obvious, and even if it exists it needs explaining.

One may argue, of course, that the dog's lack of autonomy is somehow part of the cause for the dog's lack of understanding of concepts such as property and its inability to track moral reasons. This, however, is unnecessary speculation. After all, dogs are also incapable of high aesthetic appreciation, and they cannot appreciate the wisdom of a quarterback's decisions when they watch football, either. We do not feel any particular need to say that a dog's failure to appreciate Beethoven or to judge Michigan's offensive line has to do with its lack of autonomy; and our tendency not to fault dogs for not

responding to moral reasons may be quite analogous to our tendency not to judge them critically as aesthetic philistines or as bad judges of football games. Which leads us to the point I made earlier. There are many abilities that are unique to humans or to human brains. Humans can read novels, humans can watch television, humans can use tools, humans can fall in love, humans have second order desires, humans have internal conflicts, and so on. Maybe all of these things can be traced to one property called 'autonomy', or maybe they have little in common except for requiring a high-caliber brain or the ability to reflect. Presumably some of these things have to do with moral responsibility and some do not. Which of them are relevant to moral responsibility and how they relate to each other are fascinating questions, but one cannot assume without argument that just because something is a unique property of humans, that something is the backbone of moral responsibility.

CATs and Applied Ethics

Perhaps the closest thing we have to a pre-theoretical notion of autonomy is the notion of autonomy as used in applied ethics, especially medical ethics. Talk of personal choice and of minding one's own business is central to the 'folk' value theory of the United States, and so is the idea of informed choice, of being an educated consumer, and discussions of paternalism and autonomy in medicine is to a large extent driven by intuitions about these things.

But just like the class of autonomous action stipulated by CATs often appears to be much smaller than the class of action for which agents are morally responsible, the class of autonomous actions stipulated by CATs often appears much smaller than the class of

actions that a patient, say, has the right to perform without paternalistic intervention, actions that are "one's own business". *The Field Guide to Psychiatric Assessment and treatment* (Bauer 2003) discusses the conditions under which a person in need of medical treatment – whether psychiatric or otherwise - should be regarded as competent for the purpose of medical decisions making. If a patient is judged incompetent according to these guidelines, paternalistic interventions may be indicated in her case that would not be allowed in the case of a patient who is judged competent. To establish competence, the *guide* tells us, is to establish that a patient has the following four abilities with regards to her medical needs:

- To understand the relevant facts
- To appreciate their relevance to one's own personal situation
- To rationally manipulate the information to arrive at a choice
- To communicate that choice

One striking thing about this list is that Frankfurt's or Velleman's theories of autonomy, as well as many other CATs, do not have a self-explanatory, direct connection with any of these items, and it's not clear how a clinician may find guidance in these CATs. There is nothing in the guidelines about hierarchies of mental states, alienation, a subjective sense of passivity or activity, mental conflict, or wholeheartedness. Many treatment decisions that are not autonomous by many CAT standards will be left to the discretion of the patient according to these guidelines. The paradigmatic type of persons who would be excluded by these guidelines would be the psychotic, the person who is really manic or depressed, children, the retarded and the extremely forgetful - none of which is given a

lot of attention in CATs. Unwilling addicts, compulsives, and people who are torn by conflicts often fit the list and in such cases are allowed to refuse treatment. Naturally, there could be some connections between some CATs and the competence guidelines. Perhaps, for example, a CAT that places a lot of importance on reflection and rationality may tell us something about the significance of rationally manipulating information. But again, such a claim would have to be researched, elaborated and defended. It would be substantial, not the stuff of footnotes. As in the case of explaining what exempts dog from blame, a CAT-oriented explanation of the intuitions underlying the competence guidelines may have to compete with more simple-sounding explanation, such as explanations in terms of cognitive limitations that do not speculate that something deeper, like "structure of will", needs to be behind them (I take it, if one were to claim that CATs describe essentially human characteristics and only humans have a moral status that precludes paternalism, one would be unable to get from here directly to establishing a connection between CATs and paternalism for the same reason I outlined when discussing the plurality of essentially human qualities in the previous section).

Note that there are some very important uses of the word 'autonomy' in medical ethics that are clearly not the same as the use of the word in moral psychology / agency theory. For example, one is supposed to "increase the autonomy "of a patient, or to her ability to "make autonomous decisions", by making sure not to withhold essential information from her and to provide her with additional information if she asks for it. I agree that paternalistic withholding of information from patients is generally wrong and that supplying patients with as much information as they desire is generally right. I doubt, however, that anyone wishes to claim, or that any CAT implies, that an *ill-informed* decision cannot be an instance of autonomous *agency*. Some may wish to claim that an

irrational decision cannot be an instance of autonomous agency, but being ill-informed – either because you have been deceived or simply because the relevant information is not available to you at the time when you have to decided – is not the same as being irrational. Columbus' decision to sail west may have been very uninformed, but not necessarily irrational and not at odds, for example, with the criteria for autonomy proposed by Frankfurt, Velleman, and Ekstrome. Giving a patient more information may also make her more 'autonomous' in the sense of making her less dependent on other people – the way one is more autonomous if one can fix one's own car than if one cannot. Few think that being unable to fix one's own car represent a defect of *agency* (again, Frankfurt, Ekstrom, or Velleman would not imply anything of that sort).

Implications for Complex Autonomy Theorist

Imagine that a Complex Autonomy Theorist responds to me simply by saying that her main purpose in developing her CAT has never been, primarily, to explain moral responsibility and/or to aid applied ethics, or that, having heard my arguments, she renounces any claim that her view has obvious implications for moral responsibility. Frankfurt, for one, explained in many a question period at conferences that his latest theories simply are not theories of moral responsibility. If a CAT is not meant to say anything about moral responsibility or applied ethics, if moral responsibility or applied ethics are never claimed to be the subject of the CAT, it seems as if what I say has no impact on it. It obviously does not have one type of impact. If a certain CAT never claims to be relevant to moral responsibility, it is no criticism of it that it would not serve as a good foundation for a theory of moral responsibility or of patients' rights. This would be like criticizing a metaphysician for not doing, say ethics.

But other kinds of caution are indicated if one is to develop a theory of autonomy that is not committed to any claims about moral responsibility or the limits of permissible paternalism. Moral responsibility and the limits of permissible paternalism are subjects about we have plenty of pre-theoretical intuitions, however much they may conflict with each other. They subjects about which we are forced to think fairly explicitly, even if not clearly, by personal life decisions and by judges and legislators. Divorcing discussion of autonomy from these pre-theoretical intuitions makes it more of a challenge to remain clear on the question of what exactly we are discussing and debating when we are discussing and debating autonomy. Consider, for example, the following paragraph from Frankfurt:

> Thus Agamemnon at Aulis is destroyed by an inescapable conflict between two equally defining elements of his own nature: his love for his daughter and his being devoted to the welfare of his men. When he is forced to sacrifice one of these, he is thereby able to betray himself. Rarely, if ever, do tragedies of this sort have sequels. Since the volitional unity of the tragic hero has been irreparably ruptured, there is a sense in which the person he had been no longer exists. Hence, there can be no continuation of his story. (Frankfurt 1999, 139)

The literal-minded (or, in this case literary-minded) reader may point out that there are several sequels to the tragedy of Agamemnon. He leads his men to war and victory (see Homer) and returns home, where he is killed by his wife who wishes to punish him for

sacrificing their daughter (see Aeschylus). But Frankfurt only says that there is *a sense* in which Agamemnon "no longer exists", allowing for other senses in which he keeps existing. Still, if one is interested in moral responsibility, one may reasonably ask for a to hear more about the exact sense is in which the Agamemnon who killed his daughter "no longer exists" after the killing. Does it make sense to punish the returning Agamemnon? After all, the person who decided to kill his daughter "no longer exists".

Obviously, Frankfurt does not wish to imply any such counter-intuitive claim about Agamemnon's moral responsibility, or about the moral responsibility status of anyone whose 'volitional unity has been irreparably ruptured'. But if he does not mean any such thing, what does he mean? Does Frankfurt, in saying that Agamemnon 'no longer exists', merely want to tell us that death (and subsequent replacement by a different person) is often a good *metaphor* for the state of having betrayed the object of a true love, as such an act changes the traitor deeply? We already know that, in so far as we already use such expressions as "the person I was in the sixties doesn't exist anymore" or "breaking up with her will kill me". One would assume that Frankfurt wishes to make a stronger claim than the claim that 'destruction' makes a good metaphor for Agamemnon's psychological predicament- but a stronger claim that does not have counter-intuitive implications on moral responsibility (and does not imply, for instance, that post-Iliad Agamemnon is not entitled to the property of the pre-sacrifice Agamemnon unless the latter made an appropriate will). I do not doubt that there may be such a claim, but Frankfurt does not make it easy for us to understand what claim he wishes to make.

Velleman, in "Identification and Identity", argues that Frankfurt is wrong. Agamemnon does not destroy himself as an agent. Frankfurt, in "Rationality and the Unthinkable",

seems to imply that some actions that would not be regarded as autonomous by Velleman (akratic actions that are surpassing for the actor) can be autonomous, because they follow from the agent's volitional essence (see his discussion of Lord Fawn). Whatever the precise nature of either view, it is reasonable to pose the following question:

Assume that one philosopher, such as Frankfurt, argues that Agamemnon destroys himself as an agent, while another, such as Velleman, argues that he does not. On the other hand, in the case of my akratically breaking up with a friend, Frankfurt holds that my action may be autonomous (if it came from a deep enough place in my volitional structure) while Velleman holds that it is not autonomous (because it took place without what he calls "my active participation"). Suppose further that *neither* philosopher is committed to anything about Agamemnon's, or my, status with regard to moral responsibility, or about our status as competent persons as far as medicine or the law are concerned. How am I to judge whether to prefer Velleman's theory or Frankfurt's? Frankfurt and Velleman seem to disagree. What exactly are they disagreeing about? What intuitions, exactly, are they trying to capture (so that one of them may capture them better than the other?). I have argued before (following Velleman himself, to some extent) that Frankfurt and Velleman are talking about different things, which can be called autonomy and authenticity. Ekstrom, in this volume, says that that contra Velleman and me, Frankfurt and Velleman *are* arguing about the same thing. They have conflicting substantial views of something – autonomy- and while Velleman's view implies that autonomy is somewhat similar to what we normally call self control, Frankfurt's view implies that autonomy diverges widely from the "self control" as colloquially understood and is more similar to what some romantic people would call authenticity. But I now think that if we cannot think of autonomy as related in traditional ways to moral

responsibility and the limits of paternalism, even the meta-disagreement of Ekstrom and myself does not have a clear subject matter. Does Ekstrom merely doubt the usefulness of my preferred definitional scheme, or does she join the Velleman-Frankfurt argument with a third, broader view of what autonomy is?

In other words, if one's view of autonomy is not meant to be about moral responsibility or about the limits of permissible paternalism, and so is not tied in a clear way to ignitions about moral responsibility and the limits of permissible paternalism, it seems important that one makes it clear what one's view is about, what sort of intuitions it does attempt to capture. I don't think that "intuitions about when we feel in some sense are not really ourselves" is by itself an answer, because once we agree that such statements as "it's my resentment speaking, not I" and "Agamemnon is no longer Agamemnon" are not fundamentally about moral responsibility or permissible paternalism (we already assume that they are not literal statements of personal identify), then no one tells us that all of the intuitions that express themselves in paradoxical expressions about a person not really being himself or herself are of a piece. All kinds of things can cause a person who owns a house to say "I don't really have a home", a person who is gainfully employed to say that "I don't have a real job", a person well-versed in geography to say that "Calgary is not really a city". The feeling that you don't have a home can strike you because you travel too much, because the people who live with you are so hostile that you feel more comfortable when you are away from them, because you are not emotionally attached to the place you are in, still missing another one. "I don't have a real job" can be said by a person who likes her job so much that she can't believe she is being paid for it, a person who feels that she should be making a lot more money, or a person who longs for a more stable and conventional way of life in lieu of her impossibly adventurous one. Similarly,

"this is not really me talking" and "this is not really him talking" type statements, if they are neither literal nor about moral responsibility, are just as likely to come from different intuitive sources as they are to be about one thing called autonomy.

So theories of autonomy that are not straightforwardly related to intuitions about moral responsibility and applied ethics should be clear as to what they are about. And they could well be about important things. For example, some ways of talking about how we can make our lives more autonomous appear to revive, within the analytical tradition, the search for peace of mind (arataxia) or the good life (eudemonia), or at any rate, for something that could alleviate, to some degree, our sense of being helpless before the slings and arrows of fortune and the mental turmoil they create.

References:

Bauer, Mark, Field Guide to Psychiatric Assessment and Treatment, Lippincott Williams and Wilkins, 2003

Frankfurt, H. 1999. *Autonomy, Volition and Love*. Cambridge: Cambridge University Press

Fischer, J. and Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Velleman, J. 2003 "Identification and Identity." In S. Buss and L. Overton (eds.), *Contours of Agency*

[i] The claim that we have little control over the development of our character is made and defended by Sher (2001).

[ii] Thanks to David Velleman for bringing this objection to my attention.