CURRENT MORAL AND SOCIAL ISSUES
THE COMPLETE NOTES

1. Death of the Organism vs. Death of the Person

Recall that Jeff has been saying he thinks there are really two concepts of death:

- Biological death = the irreversible cessation of the integrated functioning of the human organism as a whole.

- Personal (or biographical) death = the annihilation of the human self *via* irreversible loss of the capacity for consciousness.

It's worth noting that the definition of personal/biographical death needs to be changed just a bit if it is to suit Jeff's purposes. It's plausible that not all mental states are conscious, and it's also plausible that not all mental states had by a given subject are even *capable* of being brought to consciousness. Perhaps, for instance, some people have certain unconsciously held prejudicial beliefs that, on conscious reflection, they would vehemently claim to reject, though the beliefs would still influence their behavior when they aren't being reflective. Arguably any entity that has mental states like this should be considered a living person. But it's at least metaphysically possible that someone might lose all their conscious mental states while retaining the unconscious ones. If the initial definition of personal/biographical life is right, this shouldn't be possible. So, it might be better to state the definition as follows:

- Personal (or biographical) death = the annihilation of the human self *via* the irreversible loss of the capacity for having (conscious or unconscious) mental states.

Understood either way, biographical death comes apart from biological death. Here is a table to illustrate cases in which the two notions connect and disconnect:

	Biographical Death	*No Biographical Death*
Biological Death	Normal buried corpse	Brain transplant cases
No Biological Death	Trisha Marshall, Marion Ploch, cases of PVS, anencephalic infants	You and me

2. Why Brain Death Isn't Biographical Death or Biological Death

A major claim that has emerged from last week's lectures is that brain death is not equivalent to either biological death or biographical death. Here "brain death" means the irreversible cessation of the functioning of the *entire brain*, including both the higher brain (the cerebral hemispheres) and the lower brain (the brain stem). Let's rehearse the arguments for both halves of this claim.

There were two arguments that brain death is not equivalent to biological death. One of them is made plausible on the basis of the cases from the first chapter of Singer's Rethinking Life and Death. Recall that both Trisha Marshall and Marion Ploch were kept biologically

alive *via* respirator and feeding tube long enough to support the growth of fetuses for several months. Both, however, were entirely brain dead. So, it seems clear that brain death isn't sufficient for biological death.

The other argument can be constructed on the basis of thought experiments involving brain transplantation (or, understood a bit more accurately, full organism transplantation). If someone's brain were removed from their skull and put into another skull and appropriately hooked up to the new body, it could, at least in principle, live on. The human organism from which it was taken would, however, not live on: integrated functioning would eventually irreversibly cease. And so a biological death would have occurred. A biographical death, however, needn't have occurred: indeed, given suitably advanced technology, we can imagine that the brain that was transplanted could be kept conscious throughout the entire process. So, it seems clear that brain death isn't necessary for biological death.

Turning to the other half of the claim, the equivalence of brain death and biographical death can be undermined by reflection on cases involving persistent vegetative state (PVS). Recall that, in most cases of PVS, the higher brain irreversibly ceases to function while the lower brain does not. Since the higher brain is the seat of consciousness, the capacity for consciousness is irreversibly lost. But, since part of the brain typically lives in PVS, these are not strictly cases of brain death. Since biographical death plausibly does involve the irreversible loss of the capacity for consciousness, these are cases of biographical death without brain death. So, brain death isn't necessary for biographical death, and hence the two aren't fully equivalent. Still, the other half of the equivalence plausibly does stand, since, unless one believes in the immortality of an immaterial soul, brain death is sufficient for biographical death.

We can usefully sum all this up in the following tables:

	Brain Death	*No Brain Death*
Biological Death	Normal buried corpse, but also disembodied souls, if there are any	Brain transplant cases
No Biological Death	Trisha Marshall, Marion Ploch	You and me, but also people in a PVS and anencephalic infants

	Brain Death	*No Brain Death*
Biographical Death	Normal buried corpse, but also Trisha Marshall and Marion Ploch	PVS and anencephalic infants
No Biographical Death	Disembodied souls, if there are any	You and me, but also the brain transplant cases

3. Against the Sanctity of Life (in One Sense)?

All the arguments so far have mainly concerned metaphysical theses about the nature of death and the conditions under which it, in one sense or another, occurs. So, they strictly

speaking leave open the moral issue of whether it is ever permissible to intentionally bring about a death. Nevertheless, it is arguably easier to think about this issue once the metaphysical questions are settled. For, if we accept the idea that there are two crosscutting types of death – biological and biographical – we see that there are really two questions here:

A. Is it ever permissible to intentionally bring about a biological death?

B. Is it ever permissible to intentionally bring about a biographical death?

There is certainly some strong intuitive pull in the direction of saying 'no' to question (B) but 'yes' to question (A). Indeed, if one finds it intuitive that one ought to do the better of two actions, and that it is better to use limited resources to save biographical life rather than saving *merely* biological life, one might even venture the strong claim that it's sometimes *obligatory* to intentionally bring about biological death. Accordingly, there is some strong intuitive pull of denying the idea that an entity acquires moral status (i.e., has rights) simply in virtue of being a living being, at least in the biological sense of 'living'.

Of course, this is not beyond dispute. It does, however, increase the pressure on the traditional idea of the "sanctity of life", and puts a burden of proof on its defenders.

4. Why We Aren't Our Organisms

Another major claim from last week's lectures is that we human persons are not (numerically) identical to the biological organisms that we happen to inhabit.

Jeff gave two arguments for this claim. The first comes from the already familiar case of brain transplantation. Recall that, in this case, some person's brain is extracted from their skull and popped into the skull of a new body, and is appropriately hooked up to its nervous system. Meanwhile, the old body irreversibly cases to function as an integrated whole. But the old body was, until it stopped functioning, a full-fledged human organism in much the same way in which the bodies in the two cases from Singer's first chapter were, and in which the body of a locked-in patient receiving external support would be. Indeed, it remained the same organism before and after the brain transplant, though it lost one of its parts. If, prior to the transplant, you could have been claimed to be any human organism, it would have been this one. But you do not cease to exist after the brain transplant, while it does. So, it follows that you were never identical with it, though you did contingently inhabit it.

A second argument emerges from reflection on cases of *dicephalus*. Dicephalus occurs when a human zygote divides incompletely, resulting in the later birth of twins conjoined below the neck. It is clear that there are two distinct people in these cases: each twin has his or her own private mental life, character, sensations (though just from one side of the body), and physical control (though of just one side of the body). Nevertheless, there is only one human organism (albeit with some duplicate organs above the neck). It's a matter of logic that two distinct things can't be identical to one thing. So, it's a matter of logic that neither of the twins is identical to this organism. So, some human persons aren't human organisms.

5. Numerical vs. Qualitative Identity, the Triplets Case and What Matters

Jeff went on a bit of a tangent on Friday and talked about a further case involving triplets. The case isn't directly relevant to any of the claims he was arguing for, but it's worth rehearsing and remembering, since it does illustrate a rather striking point.

Recall that, in this case, we imagined that the brains of two of the triplets were removed and disposed of, while the two hemispheres of the remaining triplet were separated and placed within the skulls of the now brainless triplet. It is a real fact that a person can survive the destruction of one of the hemispheres of his brain and, indeed, have all his memories and intelligence left intact, as long as it's done early enough in life. Although there haven't been any actual cases of this, it's at least metaphysically possible that the remaining two triplets which now have one hemisphere could end up having exactly the same apparent memories, the same personalities, beliefs, and so on. Nevertheless, as a matter of logic, it seems very difficult to sustain the claim that the original triplet whose brain was divided survives. He clearly can't survive as both, since two distinct things can't be identical to one and the same thing. And there is no way to claim that he survived as exactly one of them, since, given what we've stipulated about the case, it would be arbitrary to choose one as his surviving self over the other: they are, after all, mentally exactly similar. So, it looks as though we must conclude that the original triplet who had his hemispheres divided ceased to exist.

Now suppose that, before having his hemispheres divided, the third triplet was asked whether he would prefer (i) to have them divided, or (ii) to have his entire brain annihilated. We know he chose (i), and I think we'd agree that his choice seems very sensible. But how *can* it be, if he ceases to exist in both cases? In reflecting on cases like this one, it has seemed to many that the only conclusion we can sensibly draw is that retaining one's numerical identity in the sense of remaining one and the same existing thing over a period of time isn't ultimately very important to us. What is important is simply *being psychologically continuous with something or other in the future, irrespective of whether that future thing is numerically the same as the thing that now exists*. This is a striking conclusion, since it certainly feels pretheoretically obvious that it really matters to us whether or not *we*, numerically speaking, will continue to exist.

6. The Morality of HESC Research

6.1. *Assisted Conception vs. HESC Research*

Before talking directly about the morality of HESC research, it is worth contrasting it with the morality of something on which a greater number of ordinary people are in agreement – viz., assisted conception. Importantly, assisted conception standardly involves the creation of more embryos than will actually be implanted, and the extra embryos are typically allowed to die. Lots of people accept the idea that assisted conception is morally permissible. One might, however, wonder why this is so, given that at least some of these people also oppose HESC research. Is there really a moral discrepancy between the two activities?

Well, there are at least two seemingly relevant differences. On the one hand, the distinction between killing and foreseeing death is in play. In HESC research, embryos are intentionally killed, whereas, in assisted conception, nothing bad is intended for the extra embryos: their deaths are merely foreseen. It does seem to be part of commonsense morality that performing an action that involves merely foreseeing a death is less objectionable that intentionally killing. Secondly, in HESC research, embryos are created *to be used merely as a*

means for an end that requires killing them, whereas, in assisted conception, the extra embryos that are created are not really used as means to any end. It is also part of commonsense morality that there is something seriously objectionable about using someone as a mere means. So, at least if we set aside the question of whether an embryo really is a *someone* worth speaking of, there is a basis for an argument for a moral discrepancy between the two cases.

For these reasons, it doesn't seem that we can simply reason as follows: (1) Allowing embryos to die in the process of assisted conception and killing embryos for HESC research are morally indistinguishable; (2) allowing embryos to die in the process of assisted conception is not morally objectionable, (3) so, killing embryos for HESC research is not morally objectionable.

Still, there does seem to remain *some* tension between the widespread acceptance of assisted conception and the much less widespread acceptance of HESC research. Although creating something and then allowing it to die is not *as bad* as creating something, using it as a mere means, and killing it, it surely isn't always entirely unobjectionable either, at least when keeping it alive is a genuine alternative, as it actually *is* in the case of assisted conception. If we really firmly believed that embryos have a moral status that is comparable to that of cognitively normal adult human persons, it is hard to see why the acceptance of assisted conception would be as widespread as it is. This, then, seems to cast into doubt whether most people really do believe that embryos have a moral status that is comparable to that of cognitively normal adult human persons.

6.2. *Monozygotic Twinning*

Jeff has given a further *prima facie* reason for doubting that most people really do believe that embryos have a moral status that is comparable to that of cognitively normal adult human persons. It is based on facts about relatively common cases of *monozygotic twinning*. In monozygotic twinning, a zygote divides to form two embryos. This usually occurs around a week or two after conception. Now, when the zygote divides, it arguably ceases to exist. For, using some reasoning that should now be familiar, there is no principled reason to think that it survives as one or the other of the two embryos. And, as a matter of logic, it can't survive as both, since the two embryos that are created are distinct. As Jeff argues, if we really thought that an embryo at any stage in its existence has a moral status that is comparable to that of cognitively normal adult human persons, we ought to regard monozygotic twinning as *tragic*, since it involves the ceasing to exist of someone who matters. And, if that's right, there would seem to be a strong *prima facie* reason to try to prevent monozygotic twinning. But this is absurd, and no one believes it. So, it certainly seems that we ought to reject the claim that an embryo at any stage in its existence has a moral status that is comparable to that of cognitively normal adult human persons.

6.3. *The Relevance of Jeff's Earlier Arguments*

Setting aside these points, a case against the impermissibility of killing embryos for stem cell research can be gleaned from some of the claims for which Jeff has already argued.

If Jeff's arguments from last week are right, we ought to reject the claim that we are identical to the biological organisms we happen to inhabit, and also reject the claim that a being without any capacity for consciousness couldn't be regarded as what we essentially are: namely, persons. Given the first point, we ought to reject that *we* ever existed as embryos in

the early stages of development at which they are to be used for HESC research. Indeed, we shouldn't believe that something like *us* can really come into existence until about twenty weeks after conception, when brain development starts noticeably occurring in the fetus. Given the second, we ought to regard the killing of such embryos as (other things being equal) comparable to the killing of a completely brain dead human organism, an anencephalic infant, or a human organism in a PVS. At any rate, the burden of proof certainly falls on someone who opposes killing embryos for HESC research but regards as permissible the killing of brain dead human beings, anencephalic infants, etc., to explain the difference.

6.4. *Why (Early) Embryos May Not Be Human Organisms*

One ostensibly significant difference that we'll come to (and reject) in a moment is the idea that the embryo has the (relatively unassisted) potential to become one of us. But first it's worth noting that *even if* you oppose Jeff's arguments that we are not the organisms that we inhabit and reject the claim that moral status turns on the capacity for consciousness, you don't *ipso facto* get any argument for the impermissibility of killing embryos for HESC research. The reason for this is that it is quite controversial to even claim that embryos in the early stages of development at which they are to be used for HESC research *really are human organisms*. Human organisms are standardly understood as entities with human genes that are composed of various living parts that function together in an integrated way to sustain a single life, and that are not themselves parts of another living biological entities. During the first two weeks after conception, however, the cells that compose the embryo do not yet serve sufficiently different functions to make it plausible to claim that they are coordinated to sustain a single life. All that exists, on any reasonable understanding, is simply a collection of nearly qualitatively identical cells living within a single membrane.

So, it is hard to see how the embryo during the first two weeks after conception can even be claimed to be a human organism. If it isn't, then even if one accepts that we *are* essentially human organisms, and that all human organisms are equally bearers of moral status, one gets no argument to oppose the killing of embryos for HESC research. For, after all, these embryos are killed within the first two weeks after conception.

6.5. *Can an Appeal to Potential Help?*

We now have a pretty strong battery of arguments against the impermissibility of killing embryos for HESC research. One last attempt to breathe some plausibility into this claim ought, however, to be considered. On this line, the fact that an embryo in the early stages at which it would be killed for use in HESC research has the *potential* to become someone like us gives it a significant moral status.

The main case against this line turns on a distinction between two different ways of understanding a potential. On the one hand, there is what we can call *identity potential*. Say that X has the identity potential to become Y iff X has the potential to become numerically identical to Y. An example of such a potential might be the potential that Obama had in 2007 to become the next President of the United States. He literally had the potential to be *numerically identical* to the next President of the United States. On the other hand, there is what we can call *nonidentity potential*. Say that X has the nonidentity potential to become Y iff X has the potential to become Y, but not by becoming numerically identical to Y. An example of such a potential is the potential of a sperm and egg to become a zygote. The

zygote clearly can't be numerically identical to the sperm, the egg, or their mereological fusion. And so the sperm and egg also can't potentially be numerically identical to the zygote. So, the only sense in which the sperm and egg have the potential to become a zygote is by ceasing to exist as sperm and egg.

Now, it is clearly the case that if X *only* has the nonidentity potential to become Y, X cannot inherit the moral status that a Y has. For, if we assumed otherwise, we would have to entertain the absurd suggestion that, prior to conception, the sperm and egg might have moral status, which they clearly don't have. But now we have a case against the attempt to oppose the permissibility of killing embryos for HESC research *if* we accept Jeff's arguments that we are not identical to human organisms. An embryo does *not*, if that conclusion is correct, have the identity potential to be one of us. So, if it has any such potential, it only has the nonidentity potential to be one of us. But, as we've already seen, it clearly can't follow from this fact that it thereby inherits our moral status.

1. When Do We Begin to Exist?

A major topic in last week's lectures was the question of when we begin to exist – when, in other words, biographical life can be claimed to begin. As you'll recall, there are at least four main candidate answers:

1. Biographical life begins at conception.

2. Biographical life begins at "viability", which is the stage at which a fetus could exist outside of the body; this period is typically taken to begin somewhere between 24 and 27 weeks.

3. Biographical life begins when the fetus acquires a living brain.

4. Biographical life begins when the fetus's brain begins to support the capacity for consciousness.

Jeff presented detailed arguments against all of these proposals except (2) and (4), and it will be useful to recall the case against each of them.

Proposal (1) is flawed for several reasons. First off, presumably the only clear rationale behind this kind of proposal would have to be the premise that biographical life and biological life coincide. We've already seen many reasons for doubting that this is true.

But, *even if* this premise were true, it doesn't actually support anything as strong as (1). For one thing, living human organisms are typically understood to be entities with human genes that are composed of various living parts that function together in an integrated way to sustain a single life, and that are not themselves parts of another living biological entities. During the first two weeks after conception, however, the cells that compose the embryo do not yet serve sufficiently different functions to make it plausible to claim that they are coordinated to sustain a single life. All that exists, on any reasonable understanding, is simply a collection of nearly qualitatively identical cells living within a single membrane. So, even if we grant that biological life and biographical life coincide, we get no argument for (1). For another thing, during the first twenty-two hours after conception begins, the genetic material from the sperm and the egg have not yet been merged; that only occurs at what is called *syngamy*. Until syngamy, it is hard to see how any distinctive new individual – whether a human organism or otherwise – could be claimed to have come into existence. So, yet again, granting that biological and biographical life coincide gives no argument for (1).

What about the second proposal? Jeff didn't say much about it, but Singer has an interesting argument against it in his book. As he points out, the claim that viability begins between 24 and 27 weeks is, if true at all, only contingently true. As he says, when it begins "depends on the medical expertise and equipment available". And largely because of this, proposal (2) seems open to embarrassing objections. Suppose that we imagine a possible world in which medical technology and expertise advance incredibly dramatically overnight, so that it would be possible to support the existence of a much younger fetus outside the womb. With this

change comes a change in *when* viability can be said to begin in this world. Now, imagine that there was a fetus that was non-viable before the technological change but viable afterwards. If we accept proposal (2), we have to claim that it went from being lifeless to being alive overnight *just because* of the technological change. This is absurd. Even more absurd is what happens when we imagine the reverse of this case. Suppose that the advanced technology and expertise is lost overnight as a result of some global disaster. And imagine a fetus that was viable with the help of this technology and expertise but not viable in its absence. If we accept proposal (2), we have to say that it went from existence to nonexistence overnight. This seems even crazier.

What about proposal (3)? Well, it seems plausible to demand as a constraint on an acceptable theory of the beginning of biographical life that it cohere with the best theory of the ending of biographical life. We've already seen that the claim that brain death and biographical death are coextensive is open to serious objections – e.g., that it implies that a person in a PVS who has altogether lost the capacity for consciousness is biographically alive. If we find that implication implausible, we ought also to reject the claim that a fetus that does not have the capacity for consciousness is biographically alive. But some fetuses lack the capacity for consciousness but have living brains – viz., anencephalic infants. So, proposal (3) seems unacceptable, too.

The failure of these proposals provides an indirect argument for proposal (4), since it is the only clear alternative left. But there is independent support for proposal (4), since it is the direct analogue of the best account of biographical death, which views it as the irreversible loss of the capacity for consciousness.

2. Abortion

3.1. *Early vs. Late Abortion and Consequences of the Preceding Discussion*

Let's distinguish between *early* and *late* abortion. An early abortion is one that occurs before a fetus's brain has developed enough to support the capacity for consciousness; a conservative estimate would have it that this is any point earlier than twenty weeks after conception. A late abortion is one that occurs after this point, and so, by parallel estimation, at twenty weeks or later.

An important thing to realize is that many of the points from our discussions of the permissibility of ending biological life, of HESC research, and of the beginning of biographical life have serious *prima facie* implications for the moral status of early abortion. If we think it is permissible to end the biological life of a human organism in a PVS *because* (i) this organism has irreversibly lost the capacity for consciousness, and (ii) moral status depends on the possession of the capacity for consciousness, it seems hard to see why we would deny that an early abortion is permissible. Of course, the fetus in an early abortion does have the potential to develop a brain that will support the capacity for consciousness. But, as we already saw earlier, it does not seem legitimate to regard this as a reason for treating it as a being with all the rights of an infant that has that capacity.

What else might create a discrepancy between the cases? Perhaps it could be insisted that a mother who voluntarily became pregnant has a special responsibility towards the early fetus, but that nothing similar can be claimed about the doctor who is considering pulling the

respirator and feeding tube from the PVS patient. Well, there does seem to be an asymmetry between the relations that the mother and the doctor bear to the respective organisms, but it isn't clear that the asymmetry really is one of responsibility for caring. Indeed, it is hard to see how a being with no capacity for consciousness could exert such a claim on the person who brings it into existence. Imagine, as a rough analogy, that some person deliberately finds a way to make a plant sprout and grow on part of his body, but then finds that living with the plant attached to him is much more burdensome than he anticipated. It seems odd to claim that he has any responsibility that prevents him from simply killing it. Why should the early fetus be different?

The only answers to this question seem to involve a retreat to the fact that the fetus is human, or that it has the potential to develop a brain with the capacity for consciousness. But we've already seen that neither mere humanity nor potential for the capacity for consciousness are morally relevant. So these answers won't be of much help.

3.2. *Marquis's Argument against the Permissibility of Early and Late Abortion*

There is a slightly different tack one could take that is represented by the position that Don Marquis defends in his "Why Abortion is Immoral", but it strikes me that it is ultimately susceptible to the same objections as the argument from potential, as well as some further objections. Marquis seems to argue as follows:

1. Killing someone is seriously wrong because it deprives that person of all the experiences, activities, projects and enjoyable things that would have otherwise been in that person's future.

2. So, any act that involves depriving an entity that has such a future is seriously wrong.

3. Both early and late abortions are examples of such acts, because "[t]he future of a standard fetus includes a set of experiences, projects, activities, and such which are identical with the futures of adult human beings and…young children".

4. Therefore, both early and late abortions are seriously wrong.

One large problem with this argument is that premise (3) seems to be false if we reject that we human persons are identical to the organism that we happen to inhabit. If an early fetus is just a human organism and not a human person, it is a mistake to say that *it* has a future that includes experiences, activities, projects, and so on, since only a being with the capacity for consciousness could have such a future. All it really has is the potential to develop an organ that would sustain such a capacity, and give rise to the existence of a person who would have such a future. So, it seems that the best Marquis can do is to retreat to a version of the argument from potential. But we've already seen that this argument fails.

Another problem with the argument is that the general idea that supports premise (1) entails very bizarre conclusions. Suppose that A and B have the same life expectancy, but that B is older than A by ten years or so. The fact that B is older does mean (other things being equal) that B has fewer experiences, activities, projects, and so on, to look forward to than A does. If we think that the basis for the wrongness of killing is the deprivation of a future

that includes experiences, activities, projects, and so on, it seems that we ought to think that killing B would be less wrong than killing A, since A's future includes more of these things than B's. But this is absurd. If we reject that idea that an action's degree of wrongness covaries with the amount of good of which it deprives a person, it is hard to see what the motivation is for accepting (1). We ought, in other words, to prefer some other account of the features that make killing wrong.

A final problem with this argument is that it fails to distinguish between *prima facie* and *ultima facie* wrongness. An act is *prima facie* wrong when there is some moral reason – strong or weak – not to perform it. An act is *ultima facie* wrong when all the moral reasons that bear on performing the act count against performing it. Plausibly, how we ought to act turns on the balance of all the reasons in play, not just on one of them, and hence on *ultima facie* rather than *prima facie* wrongness. But Marquis has left it an open question whether the mother's right to choose how her own life is to proceed is a further reason that counts in favor of at least some abortions, and, if it is, how weighty a reason it might be. And it is at least dialectically coherent for Marquis's opponent to insist that this reason might outweigh whatever moral reason there is not to kill the fetus. Perhaps he thinks it is obvious that this latter reason is weightier, but he does nothing in the paper to argue for this conclusion. Of course, at times, he does seem to suggest that he's only arguing for the *prima facie* wrongness of abortion. But since the real question is how we ought to act, he has an obligation to say something about *ultima facie* wrongness, too, since that and not mere *prima facie* wrongness is what determines how we ought to act in these cases.

1. The Naïve Argument and Jeff's Objections

We can view the main cases we've seen against the impermissibility of abortion as responses to the following argument:

The Naïve Argument for Impermissibility

1. Every innocent person has the right to life. (Assumption)
2. Every fetus is an innocent person. (Assumption)
3. So, every fetus has the right to life. (Follows from 1 & 2)
4. If X has the right to life, then causing X to die violates X's rights. (Assumption)
5. So, causing a fetus to die by aborting it violates its rights. (Follows from 3 & 4)
6. Violating an innocent person's rights is always impermissible. (Assumption)

---

7. So, abortion is always impermissible. (Follows from 5 & 6)

Before we started discussing Thomson, most of our negative discussion targeted premises (1), (2) and (3). In particular, Jeff offered the following two criticisms that jointly take down (1 – 3):

- A. *Against (2) and (3): Biographical vs. Biological Life.* There are strong reasons for thinking that biographical life only begins when the capacity for consciousness is acquired. But, prior to 20 weeks after its conception, it isn't plausible that a fetus has the capacity for consciousness. So, there seems to be no reason to accept (2). We've also seen that there are strong reasons for thinking that a being has moral status only if it has the capacity for consciousness; intuitively, PVS patients, anencephalic infants, et al., do not have the kinds of rights that you and I have. So, there also seems to be no reason to accept (3).

- B. *Against (1) and (3): Gradualism about Moral Status.* It is arguable that the acquisition of moral status is a *gradual* matter, and that some entity does not gain all the rights that you and I have *simply* because it acquires the capacity for consciousness. After all, animals have the capacity for consciousness, but lots of people do not find it obvious that they have the moral status that you and I have. If, however, we do grant that any being with the capacity for consciousness is a person, then it will be far from obvious that (1) is true, since the right to life *might* be something that is acquired later after (though only after) the capacity for consciousness is acquired – say, once self-awareness begins.

Criticism (A) also undermines Marquis's argument. Only a being that *already* possesses the capacity for consciousness can be claimed *to lose its future as a conscious being* rather than to simply *be prevented from being able to develop into an organism that would give rise to the sort of conscious being that would have such a future.* Crucially, only the first of these is morally tragic. In the majority of cases, killing a fetus arguably only prevents a person with a future like you and me from existing.

2. Thomson's Argument

As we saw on Friday, Thomson has a more radical way of objecting to the Naïve Argument. She attacks (4) and (5) with a series of arguments from analogy. Her first argument turns on:

> *Violinist I.* You've been kidnapped and connected to an unconscious famous violinist, who will die unless he uses your kidneys for the next nine months.

As she suggests, it would be absurd if the director of the hospital said to you: "We're sorry that the Society of Music Lovers did this to you, but we cannot unplug you. After all, this violinist is a person with the right to life, and to unplug him from you would cause him to die, and that would violate his right to life, which is impermissible." But the director of the hospital would be right about *something* here: the violinist is indeed a person with the right to life. This suggests that it can't be generally true that if X has the right to life, causing X to die violates his rights. The reason why (4) is false in this case is that the fact that the violinist has a right to life *does not give him the right to use your body as a means for life support.* Since he doesn't have that *further* right, disconnecting him wouldn't violate any right of his. Thomson suggests that abortions in cases where the pregnancy is due to rape should be viewed as analogous, and hence that (5) is false.

This argument says nothing about cases where pregnancy is not due to rape. Thomson does, however, offer further reasons for thinking that we should reject (5) in other cases. The first alternative case she considers is one in which the pregnancy threatens the mother's life. She develops the following variation on Violinist I to strengthen her argument by analogy:

> *Violinist II.* Like Violinist I, except that the violinist's use of your kidneys will put such a strain on you that you will probably die.

Now, even if we imagine that, prior to learning that the violinist's use of your kidneys will put such a strain on you that you will probably die, you *consented* to the operation, it seems you would be morally permitted to back out of the operation as soon as you learn this fact. So, Thomson again reasons by analogy that not only in cases of pregnancy by rape, but also in cases where pregnancy threatens the mother's life, is she permitted to abort the fetus.

Notably, nothing yet follows about whether a *third party* would be permitted to abort the fetus for the woman. For, in general, if X is permitted to do A, it doesn't follow that any arbitrary Y is permitted to do A for X. But Thomson insists that the fact that the woman's body is *hers* would enable a third party to perform the abortion for her. To support this view, she considers:

> *Stolen Coat.* Jones steals a coat from Smith to prevent himself from freezing to death. Smith will also freeze to death if he doesn't get the coat back.

It seems clear that it would be permissible for a third party to take Smith's coat back from Jones in *Stolen Coat*, even though this would lead to Jones's death. Intuitively, this is because the coat belongs to Smith. So, by analogy, since the woman's body belongs to her, a third party could abort the fetus, which is using the mother's body for life support, and which threatens her life.

What about cases where the mother's life is *not* threatened? Thomson suggests that even in these cases, abortions may be permissible *when* they would impose a substantial burden on the mother, and when it's clear that "minimal decency" would not require the mother to

carry the fetus to term. In support of this conclusion, Thomson first notes that we should reject this argument:

i. A has a right to life.
ii. Using X is the only way to save A's life.
---
iii. Therefore, A has a right to use X.

To show that this argument is invalid, Thomson appeals to the following case:

> *Henry Fonda's Cool Hand I.* The only way that you could be saved from dying from your sickness is by having Henry Fonda's cool hand touch your fevered brow. But he is thousands of miles away, and it would burden him to travel to you.

It's clearly false that you have the right to the touch of Henry Fonda's cool hand in this case. But it is still true that you have the right to life, and that being given the touch of Henry Fonda's cool hand is the only thing that could save your life. So, (i) and (ii) can be true while (iii) is false. Indeed, if we imagine that, in *Henry Fonda's Cool Hand II*, Henry Fonda is not thousands of miles away, but just across the room, it *still* doesn't seem that you'd have a *right* to the use of his hand, though he *ought* to give it to you, and though he'd be a really crappy person not to do so. So, Thomson insists that we should also reject the following piece of reasoning:

a. A ought to give X to B, since X will save B's life.
---
b. B has the right to be given X by A, since X will save B's life.

So, by analogy, Thomson suggests that if carrying a fetus to term would impose a substantial burden on the mother, it may be permissible for her to abort it. And she also suggests that although, if carrying the fetus to term wouldn't impose any burden, it may be impermissible to abort the fetus, this is *not* due to the (putative) fact that the fetus *has the right to the use of her body*.

Now, when exactly *does* Thomson think that abortion isn't permissible? It is, I think, fair to view her as conceding that if the mother becomes pregnant *via* voluntary sex in full knowledge of a significant likelihood that it would result in pregnancy, and if carrying the fetus to term would not impose any substantial burden on the mother, it is impermissible to abort the fetus. Here the condition that it must be known to be *likely* that the sex act will result in pregnancy is important. Thomson argues for it by analogy with this exceedingly whimsical case:

> *People Seeds.* People seeds are drifting around that grow in carpets and upholstery. You install mesh screens in your windows to prevent them from drifting into your house. But, against the odds, one drifts in and takes root.

Thomson takes it that even though what resulted in the person-seed's getting into your house was a voluntary act of yours, it doesn't follow that you have any obligation to allow

the seed to grow. This is because you took precautions to vastly reduce the probability that this would happen. By analogy, if a pregnancy happens in spite of the use of effective contraceptives, it doesn't follow that the mother has an obligation to carry the fetus to term.

Her ultimate conclusion is that abortion is sometimes permissible and sometimes not. What distinguishes between the cases is whether it would be "minimally decent" of the mother to allow her body to be used as a means of life support. In cases of rape, threat or burden to the mother, or improbable unwanted pregnancy, minimal decency does not require this. Even when minimal decency does require an abortion not to be performed, this is not, or at least not principally, due to the fact that the fetus has a right to the use of the mother's body.

3. Objections to Thomson

One objection to Thomson's argument is known as the *Responsibility Objection*. According to this objection, in all cases other than cases of pregnancy by rape, the mother is responsible for the existence of the fetus. But, in many of Thomson's cases, it is not plausible that the person who kills is *responsible for the other person's need for aid*. This is very clear in both Violinist I and Violinist II. So, given the disanalogy, why think that conclusions from her cases can be extended?

Thomson's only reply to this objection turned crucially on the thought that, in cases where an effective contraceptive is used, the fact that the resulting pregnancy was unlikely frees the mother from being responsible in any *relevant sense* for the existence of the fetus or its need for aid. But people in the literature object to her appeal to improbability by noting our intuitions about the following kind of case:

> *The Cautious Hunter*. A hunter takes every reasonable precaution to avoid shooting innocent bystanders, and the objective probability of his hitting one is in fact extremely low. But the improbable does happen, and the hunter ends up shooting an innocent bystander by accident. The bystander needs a blood transfusion to survive, and the hunter has the right blood type.

Here, in spite of the improbability of the accident, the hunter does seem to have a duty to provide the transfusion, and so is responsible for the bystander's need for aid.

But this quip is too quick, since there is a distinction between *being responsible for someone's existence (or continued existence)*, and *being responsible for a need for aid that inevitably accompanies their existence (or continued existence)*. To see this, consider the difference between these cases:

> *Imperfect Drug*. A famous violinist has contracted a rare disease. The only cure for the disease is a pill that has an unfortunate side-effect: ten years after taking the pill, the violinist will likely end up with a kidney ailment which could be cured by your hooking him up to your kidneys for several months. You give the violinist the pill.
>
> *Malpractice*. Like *Imperfect Drug*, except that there are two pills, only one of which has the bad side-effect. You give the violinist the pill with the bad side-effect.

While you are responsible for the violinist's continued existence in both cases, you only have the later responsibility to hook the violinist up to your kidneys in *Malpractice*. So, if you cause

someone to exist or continue to exist, and the only way to do this also makes them require aid, you do not thereby acquire the responsibility to provide that aid.

Another objection is the *Parental Bond Objection*. In none of Thomson's cases is there a biological relationship between the person in need and the person who could provide the aid. But surely there's some intuitive plausibility to the idea that the fact that the woman is the fetus's *mother* gives her a special reason to attend to its need for aid. Thomson's only reply to this objection is a flat-footed denial of the intuition, and this, to say the least, is extremely unsatisfactory.

A final objection is the *Killing vs. Letting Die Objection*. In Thomson's key cases, the candidate provider of aid only *lets the person in need of aid allow to die*. But, according to the objection, the fetus is killed in abortions. And there is a moral difference between killing and letting die.

Thomson does reply to this objection with her Growing Child case, but a much simpler response is that it is just false that all abortions require the fetus to be killed. This is, for instance, not true of hysterotomy abortions. So, at best, the objection simply shows that abortive practices should be changed, not that abortion is *per se* impermissible.

4. Marquis's "Future Like Ours" Argument

Marquis argues as follows:

5. Killing someone is seriously wrong because it deprives that person of all the experiences, activities, projects and enjoyable things that would have otherwise been in that person's future.

6. So, any act that involves depriving an entity that has such a future is seriously wrong.

7. Both early and late abortions are examples of such acts, because "[t]he future of a standard fetus includes a set of experiences, projects, activities, and such which are identical with the futures of adult human beings and…young children".

8. Therefore, both early and late abortions are seriously wrong.

One large problem with this argument is that premise (3) seems to be false if we reject that we human persons are identical to the organism that we happen to inhabit. If an early fetus is just a human organism and not a human person, it is a mistake to say that *it* has a future that includes experiences, activities, projects, and so on, since only a being with the capacity for consciousness could have such a future. All it really has is the potential to develop an organ that would sustain such a capacity, and give rise to the existence of a person who would have such a future. So, it seems that the best Marquis can do is to retreat to a version of the argument from potential. But we've already seen that this argument fails.

A further problem with Marquis's argument arises if we grant him the claim that a fetus at any stage of its existence is an entity with a biographical life like yours and mine. Notice that his argument is only plausible if we accept in general that the badness of X's death for X is

proportional to the amount of good that would be deprived from X in dying. When conjoined with the assumption that the fetus has a biographical life, this claim implies that it should be more seriously wrong to kill a fetus (that could otherwise live a long life) than to kill a 20 year-old adult, and that it is worse for a fetus (that could otherwise live a long life) to die than for a 20 year-old adult to die. But these claims do not seem at all plausible. Indeed, the second claim implies, as Jeff pointed out, that we ought to regard the fact that 60+% of pregnancies end in spontaneous abortions as being seriously tragic – indeed, more tragic than some scenario in which a fatal disease wipes out the same number of adults who would otherwise have lived on. But we shouldn't regard spontaneous abortion as seriously tragic, and don't. So, something seems to be seriously wrong with the theory of the badness of death and wrongness of killing that underlies premise (1) in Marquis's argument.

What could replace this account? One suggestion endorsed (with qualifications to which I'll turn in a moment) by McMahan is that we adopt a view on which the badness of X's death for X is a function of two factors: (1) the good experiences (etc.) that would be lost by X's dying, and (2) X's degree of psychological continuity and connectedness with the future selves that would enjoy these goods. In particular, the idea behind this account is that the badness of X's death for X at some time *t* is computed by scaling factor (1) by factor (2), so that, if X's future self-stages are not very well continuous or connected with X's current stage, and only these future self-stages would enjoy a lot of good experiences (etc.), it wouldn't be all that bad for X at *t*.

This kind of theory is motivated by a point that we discussed a few meetings ago – namely, the fact that qualitative identity matters more to us than numerical identity. I can usefully illustrate why this theory seems to be an improvement over the simpler theory that Marquis assumes with the following scenario:

> *The Choice*. You have been kidnapped by a group of kooky neurosurgeons. They present you with two options. On the one hand, by tinkering with your brain, they could erase all your memories, replace them with false ones and give you a whole new set of desires, intentions and beliefs, making you a qualitatively very different person than you currently are. They would do this, however, without rendering you unconscious: you would just experience a very, very jarring change in your mental life, and end up feeling like a new person. If they do this, they promise to compensate you by giving your future self enough money to retire young and have a very enjoyable life. On the other hand, they could split the hemispheres of your brain, install them in new bodies, and ensure that both brain-halves would be functionally equivalent to your current whole brain, and have all the same memories, beliefs, desires, and so on. They will not, however, compensate you if they do this, and the two entities that go on living will have lives of the same quality as the life you would have had if they had never kidnapped you.

My own intuitions tell me that it would be better to choose the second option over the first. The reason is that the fact that you have so little psychological continuity with the person that exists after they tinker with your brain gives you very little or perhaps no reason to have any vested interest in his admittedly very enjoyable future. Because, however, you would have perfect psychological continuity with both of the hemisphere twins that end up existing in the second case, you have very good reason to have a vested interest in their futures,

which will be quite decent, though not as cushy as the one that the profoundly changed version of you would have in the first case. This is all in spite of the fact that the good in store for your profoundly altered self will be vastly greater than the good that is in store for the hemisphere twins.

If my intuitions about this case were generally shared, that would count strongly in favor of the account that McMahan suggests as a replacement. For my intuitions about the case follow the pattern it predicts: the good experiences that my profoundly altered future self would have in the first case are greatly discounted by the fact that my current self would have virtually no psychological continuity with this future self. And so, even though there is more good in that future, I still have more reason to prefer the scenario in which my hemisphere twins live a life just like the one I would have lived had I not been kidnapped.

Notably, this revised account yields very different predictions about the morality of abortion from Marquis's account. In particular, since a third-trimester fetus has very little psychological continuity and connectedness with its adult self, it has very little reason to care for its own sake about the good that is in store for that self. So, although it may be deprived of a lot of good by being killed, this must be discounted in the way that we've already seen. So, the revised account does not have the bad consequence that the death of a fetus is worse than the death of a normal 20 year-old who will have a long and good life. Indeed, if it's granted that the psychological continuity and connectedness of the infant is extremely weak, the fact that some mothers would avoid a lot of misfortune by aborting it may outweigh the reasons against killing it. This, in effect, is McMahan's take on late abortion: killing a third-trimester fetus is more seriously wrong than killing a fetus that's less than 20 weeks old, but it is far less seriously wrong than killing an adult, and hence may be, on balance, justified.

Now, one *does* have to be careful here if one wants to avoid the conclusion that it is not seriously wrong to kill a cognitively impaired person, a person with Alzheimer's disease, etc. What Jeff suggests we do to avoid this implication is to adopt a two-tiered theory on which different conditions explain the wrongness of killing autonomous beings with self-consciousness and non-autonomous beings without self-consciousness. The former beings are subject to a "morality of respect", while the latter are not.

1. More on the Argument from Potential

An argument that is close in spirit to Marquis's, but which is problematic for interestingly different reasons, is the following *Argument from Potential*:

1. A fetus has the potential to be a human person like you and me.

2. It is seriously wrong to kill anything that has the potential to be a human person like you and me.

3. Therefore, it is seriously wrong to kill a fetus.

As Jeff pointed out last Friday (and as I discussed a bit a few times ago), the first premise of this argument is ambiguous. For there are two very different kinds of potential:

- *Nonidentity Potential.* If an X only has the *nonidentity potential* to be a Y, an X cannot literally become *numerically identical* to a Y. Three potential examples are: (i) a lectern's potential to become a pile of sawdust, (ii) a sperm-egg pair's potential to become an embryo, and, more controversially, (iii) a lump of bronze's potential to become a statue.

- *Identity-Preserving Potential.* If an X has the *identity-preserving potential* to be a Y, an X can literally become numerically identical to a Y. An example of this is Prince Charles's potential to become King of England.

It seems clear that nonidentity potential is not morally significant, and that if "potential" in (2) is understood as "nonidentity potential", (2) is false. One reason for this is that it is obvious that contraception is not seriously wrong, or wrong at all.

Some may find this a bit quick. They might insist that the intuition that it's not wrong to destroy a sperm-egg pair or an early embryo is not a *relevant* intuition. One way someone might try to do this is by arguing that the probability that either would actually develop into a person like you and me is low, and that *this* is why it isn't seriously wrong to destroy either. Since the same point doesn't apply to all cases of potential, it might seem that we've generalized from bad examples.

But this doesn't seem to me to be an adequate explanation. Suppose that, in the remote future, someone created a machine that could fertilize an arbitrarily large number of eggs, and implant the resulting embryos into artificial wombs. Suppose that the machine could reliably do this with perfect success, and that an indefinite number of artificial wombs were made available to it for the purposes of implantation; the technology is so effective that the probability that an embryo would develop into a perfectly healthy fetus that could survive past the third-trimester is 100%. Finally, suppose that some person – say, Jones – obsessed with creating more human life gets a hold of this machine, and tells it to fertilize millions of eggs and implant the resulting embryos into artificial wombs. The probability that these would be implanted in the artificial wombs and develop into healthy fetuses that could

survive beyond the third-trimester is 100%. But although there is maximum probability that the fertilized eggs will (if the machine isn't interfered with) develop into people like you and me, it still doesn't seem like it would be wrong to stop the machine from implanting them, or to destroy an artificial womb just after an embryo has been implanted. So, seems like the argument that nonidentity potential is morally insignificant stands.

But now there is a big problem for the Argument from Potential. Jeff has argued at length that we are not our organisms. Before 20 weeks, a fetus is just an organism, not a person. As such, it only has the identity-preserving potential to develop into an adult organism. But that organism is not *by itself* morally significant: the person it houses is what is morally significant, and the fetus before 20 weeks does *not* have the identity-preserving potential to become this person. So, if we read "potential" in the argument as "identity-preserving potential", premise (1) is false. So, either way of reading "potential" leads to a false premise in the argument.

Jeff made this point in a slightly different form on Friday. As he noted, if one wants to argue that there is something bad about destroying something that only has the potential to become someone like you and me, it has to be made plausible that it brings some good into the world by realizing or having this potential. There are three ways in which this might be understood:

A. The realization of the potential by the thing that has it might be good for it.

B. There are instrumental reasons for having something around that has this potential.

C. The mere fact that X has the potential to become Y might confer special status on X.

Jeff then reasoned as follows. (A) does not seem to apply to any case in which the fetus is less than 20 weeks old, because the fetus at this stage is just an organism, and realizing the potential to become a person cannot be good for a mere organism in any relevant sense. (B) applies equally well to a sperm-egg pair, and so cannot provide the basis for any argument from potential that wouldn't overgeneralize to the absurd conclusion that contraception is morally wrong. And (C) also does not apply to any case in which the fetus is less than 20 weeks old, because the fetus only has the identity-preserving potential to become an adult human organism, and an adult human organism *by itself* just doesn't seem to be something with high moral status; after all, a PVS patient, one of Singer's brain dead patients, and an anencephalic infant are all human organisms, but they definitely don't have the kind of moral status that you and I have. So, it seems that we can argue by exhaustion that an appeal to potential is not going to help argue for the impermissibility of any abortion before 20 weeks.

2. Contingent Identity and Jeff's Objection to the Argument from Potential

There is one important way in which some people might want to undermine Jeff's objections to the argument from potential. Notice that there is a big difference between an early fetus's potential to become an adult human being and a lectern's potential to become a pile of sawdust. The early fetus does not cease to exist when it becomes an adult human being, whereas a lectern does cease to exist when it becomes a pile of sawdust. In this way, an early

fetus's potential to become an adult human being is much more like a lump of bronze's potential to become a statue: when the lump is shaped into a statue, it doesn't cease to exist.

This asymmetry opens up a line of reply for the defender of the argument from potential. Some people think (indeed, a majority of people used to think) that there is such a thing as *contingent identity*: two things can be numerically identical at one time but be numerically distinct at another time.[1] If you believe in contingent identity, then it should not strike you as an argument for the distinctness of two entities at all times to be told that the entities have different persistence conditions. If, in particular, you believe in contingent identity, you will not think that, when the statue exists, the lump of bronze is distinct from it *merely because* the lump *could* continue to exist while the statue goes out of existence. Why not? Because you'll say that when this happens, the lump *loses the property of being identical to the statue*, a property that it contingently had before the statue was hammered into an amorphous hunk.

If this is right, you will *also* not believe that the lump's potential to become a statue really is an example of nonidentity potential. But if it isn't, then if the fetus's potential to become a

---

[1] What motivates contingent identity? One motivation is that the competing view on which all facts of numerical identity are necessary leads to paradox. Here is a paradox-generating argument that illustrates this:

> 1. If you chopped off my arms, my body would continue to exist.
>
> 2. So, given (1), the persistence conditions for bodies and bodies-with-arms are different: it is possible for a body-with-arms to cease to exist while a body continues to exist.
>
> 3. If X and Y have different persistence conditions, X and Y can never be numerically identical.
>
> 4. Therefore, the whole body with arms that stood before you last Friday in Section 4 can never be numerically identical with my body.

(4) is crazy. So, something is wrong with this argument. The truth is that it is extremely puzzling to see what premise besides (3) could possibly be false in this argument. So, necessary numerical identity leads to a paradox. Contingent identity, on the other hand, affords a nice explanation of this case: the whole body with arms that stood before you last Friday in Section 4 is contingently identical to my body. The two would cease to be identical if my arms were chopped off. But, thankfully, my arms haven't been chopped off, so they're contingently identical.

Notice that we haven't made any move in the paradoxical argument that Jeff hasn't already made in a parallel argument. Indeed, his argument from brain transplantation has exactly the same form:

> 1*. If my brain were removed from my body and transplanted into a new body while my previous body was allowed to decay, I would continue to exist.
>
> 2*. So, given (1), the persistence conditions for persons and the organisms they inhabit are different: the person could exist while the organism ceases to exist. (We also know that the person could cease to exist while the organism continues to exist, given PVS cases.)
>
> 3*. If X and Y have different persistence conditions, X and Y can never be numerically identical.
>
> 4*. Therefore, from (2) and (3), persons and organisms can never be numerically identical.

Barring a sophisticated metaphysical theory, it seems that one can't agree that Jeff's (1* - 4*) argument is sound without also agreeing that the paradoxical (1 - 4) argument is sound. There are sophisticated metaphysical theories that reveal disanalogies and save necessary identity from this problem (the "problem of temporary intrinsics"), but Jeff hasn't told you about them, and he doesn't endorse them in print.

human being is analogous, it also isn't an example of nonidentity potential. But, crucially, Jeff has only given us an argument that nonidentity potential is morally insignificant. He has not given us an argument that identity-preserving potential is morally insignificant. Without such an argument, it seems that nothing blocks the original Argument from Potential from succeeding. Indeed, the points about the obvious permissibility of contraception become irrelevant, since the sperm-egg pair's potential to become a fetus is *not* analogous to the lump of bronze's potential to become the statue: the sperm-egg pair really *does* go out of existence when it develops into an embryo, and then a fetus. So, in short, the reasons that contingent identity gives us for thinking that the fetus might have the identity-preserving potential to become a human person like you and me do not overgeneralize into reasons for thinking that a sperm-egg pair also has this sort of potential. Given this asymmetry, no objection from contraception arises.

This point does have some limitations. It still seems clear that spontaneous abortions are not as bad as the deaths of human adults from pandemics. If this is right, then *while* the fact that the fetus would, on this view, have the identity-preserving potential to become a human person like you and me might give it *some* status, it does not seem to give it the same status as adult human persons like you and me. So, it seems that there will remain a lingering moral asymmetry that could block the full force of the conclusion of the Argument from Potential.

3. Identity-Preserving Potential, 20+ Week-Old Fetuses and Infanticide

It is also, however, important to realize that if identity-preserving potential does confer *more* moral status than nonidentity potential, there are important consequences for abortions beyond 20 weeks and for infanticide *even* for those who reject contingent identity. Once a fetus acquires the capacity for consciousness at 20 weeks, virtually everyone, including people who reject contingent identity, should concede that it has the identity-preserving potential to become a person like you and me. Moreover, virtually everyone should also agree that an infant has the identity-preserving potential to become someone like you and me. For this reason, one could still use the following *Restricted Argument from Potential*:

4. 20+ week-old fetuses and infants have the identity-preserving potential to become human persons like you and me.

5. It is bad to kill anything that has the identity-preserving potential to become a human person like you and me.

6. Therefore, it is bad to kill 20+ week-old fetuses and infants.

This argument doesn't tell us *how* bad it is to kill 20+ week-old fetuses and infants. But it does tell us that it is worse than contraception. And that is a nontrivial result.

Another upshot of this point is that Jeff's theory that the wrongness of killing a late-term fetus or an infant is proportional to the loss for it multiplied by the degree of psychological continuity with the futures selves whose lives are lost cannot apply so straightforwardly. For perhaps the identity-preserving potential to become a person puts something within the realm of respect, so that it is indeed seriously wrong to kill a late-term fetus or early-term infant. After all, Jeff does want to grant that people suffering from dementia are still within

the realm of respect, but there isn't as big a mental difference between them and late-term fetuses and infants as there is between late-term fetuses and infants and normal adult persons like you and me.

4. Personal Good, the Satisfaction of Preference, and Euthanasia

Another topic worth discussing briefly is Singer's view that what is good for a person is exhausted by the satisfaction of his or her preferences. In particular, I want to illustrate how the most important style of objection to this theory leads to a puzzle about the distinction between voluntary, non-voluntary and involuntary euthanasia (defined below), and the morality of each.

Notice that, according to the preference theory of personal good, if a person prefers that one state of affairs X obtains rather than another Y, then it is always better for him if X obtains rather than Y. This leads to some very bizarre implications in cases where people's preferences are irrational. Suppose that Jones irrationally desires to be in a state of intense agony. Singer's preference theory of personal good entails that the best thing for Jones is for him to be in intense agony. If we combine this theory of personal good with a preference utilitarian theory of how we ought to act (a theory on which one ought to A iff A-ing would lead to the greatest ratio of good over bad for everyone), it also follows that we ought to cause Jones to be in intense agony. But this seems wrong. Jones's desire is bonkers: he doesn't know what is good for him.

There is a way of revising the theory to avoid this result, but it has a different problem. A preference theorist might say that what is good for a person is exhausted by the satisfaction of the desires that he would have if he were fully rational. This theory may not have the bad consequence that we just imagined. But the theory seems circular, since one might think that we ought to analyze rational preference in terms of what is good for someone, rather than *vice versa*.

But let's set aside this worry about the potential circularity of the revised view. What is interesting is that, circular or not, it does seem *true* that satisfying someone's desires is good for him only if his desires conform with principles of rationality. This kind of theory has an interesting consequence for how we should think about the distinction between voluntary, non-voluntary, and involuntary euthanasia. Recall that these are defined as follows:

> *Voluntary Euthanasia*: X expresses a persistent desire to be euthanized.
>
> *Non-voluntary Euthanasia*: X is incapable of expressing any desire to be euthanized, and did not express any prior preference when he his expressive capacities were still intact.
>
> *Involuntary Euthanasia*: X expresses a persistent desire not to be euthanized.

If it is just false that what is good for someone corresponds to what would satisfy his desires, then one might wonder why we should make a moral distinction between voluntary, non-voluntary, and involuntary euthanasia in cases where a person's preferences are not in conformity with norms of rationality. If someone irrationally wants not to be euthanized, and it would really be better for him if he were, what reason could there be to honor his wish? After all, whatever else you might say, it does seem clear that we ought to do what is

best for people, and this is a case where what is best for him clashes with his desires, so that it would also seem clear that we ought to do what would frustrate his desires.

Does anyone have thoughts? I'm not claiming that there are no distinctions, but that it is puzzling how there *could* be any if we think that only rational desires are worthy of respect.

5. McMahan on Infanticide

Following Carlotta, we can state Jeff's conditional argument that infanticide should be sometimes permissible if late abortions are sometimes permissible as follows:

> *The Consistency Argument*
>
> 7. Any X's moral status is entirely a function of X's intrinsic properties. (Assumption)
>
> 8. In general, X's birth does not relevantly change any of X's intrinsic properties. (Assumption)
>
> 9. Viable fetuses would be newborn infants if they were delivered prematurely. (Definitional truth)
>
> 10. So, viable fetuses' moral status should be the same as the moral status of newborn infants of the same age, measured from conception. (Follows from 7 – 9)
>
> 11. Whether it is permissible to kill something is entirely a function of its moral status. (Assumption)
>
> 12. In some cases, it is permissible to kill viable fetuses. (Intuition)
>
> 13. So, in relevantly similar cases, it is permissible to kill newborn infants. (Follows from 10 – 12)

Exactly how forceful and plausible this argument is depends on the details of the cases in which we ought to regard the abortion of a viable fetus as permissible. Jeff focuses on two cases. The first is one many regard as a clear example of a permissible abortion, while the second is a more controversial case that Jeff thinks should be as permissible as the first:

> *Selfish Abortion.* A woman who became voluntarily pregnant learns 6.5 months after conception that if she carries her viable healthy fetus to term, she will end up suffering moderate chronic pain for the rest of her life. Extracting it alive would be disfiguring and riskier for her than having an abortion. She gets an abortion.
>
> *Altruistic Abortion.* A woman who became voluntarily pregnant learns 6.5 months after conception that her existing three year-old will die within days without an organ transplant. The needed organ is a vital one, and there are no suitable potential donors except for her fetus. As it happens, there are three other children in the hospital who could also receive transplants from this fetus, and who will otherwise die, since suitable further donors are not to be found.

There are two *prima facie* compelling ways in which someone might try to ground a moral asymmetry between these cases. One appeals to the apparent fact that the woman in *Selfish*

Abortion could be seen as having a justification from self-defense, and that this is the best explanation for why the abortion is permissible; since the same can't be said about *Altruistic Abortion*, it seems coherent to insist that it is still impermissible. Jeff rejects this explanation, because he thinks that one can only have a justification for self-defense against a *responsible threat*, and the fetus is not a responsible threat.

Another explanation appeals to the idea that one shouldn't engage in *harmful using*. Since there is harmful using in *Altruistic Abortion* but no harmful using in *Selfish Abortion*, it might seem that there should be a corresponding moral asymmetry between the cases. Jeff rejects this explanation. He notes that most people don't think the constraint against harmful using applies to higher animals in cases that are otherwise relevantly similar to *Altruistic Abortion*, and notes that the only ostensibly decent reason people give for this is that there are mental differences between higher animals and normal adult people like you and me. But, as he suggests, there don't seem to be similarly strong mental differences between viable fetuses and animals. So, if the constraint doesn't apply to animals, it also shouldn't apply to viable fetuses. (*Nota bene*: one might take this as a refutation of the idea that the constraint doesn't apply to higher animals!)

Having concluded that there is no convincing reason to deny that *Altruistic Abortion* is permissible if it is granted that *Selfish Abortion* is permissible, Jeff applies the Consistency Argument to yield the conclusion that the following two cases of infanticide are permissible:

> *Contagious Newborn.* A few weeks after a woman gives birth to a premature baby, it is bitten by a dangerous insect that infects it with a highly contagious virus. The virus poses little threat to infants but substantial threat to adults. The mother will be infected unless she leaves the infant outside in the wintry surroundings in which she is isolated; she cannot leave by herself, and it would take two weeks for help to come if she signaled for it, at which point she would be irreversibly infected. She decides to put the infant outside and allow it to die.
>
> *Healthy Newborn.* A woman dies in childbirth leaving a premature but healthy infant. The infant's biological father died months ago. There are no surviving relatives to care for the child, and the parents were so reclusive that they never had friends who came to know the infant. There are three other children in the same hospital who need organ transplants to survive. Since their organs have been impaired by illness, the doctors cannot wait for one to die and use his organs to save the others. The infant is the only available candidate donor. The doctors kill it and transplant its organs, saving the other three children.

Since, Jeff claims, *Contagious Newborn* and *Healthy Newborn* are similar in all relevant respects to *Selfish Abortion* and *Altruistic Abortion*, these are cases to which his argument applies

How might one object to the Consistency Argument? The main objection Jeff considers is a challenge to (11) from Thomson's Argument. Recall that part of what makes Thomson's argument for the permissibility of abortion work is the assumption that, since the woman owns her body and the fetus does not, it does not have a right to use her as a means of life support, and she has no obligation to allow it to do so in cases where the pregnancy is the result of rape, is threatening to her, or where giving birth would create a substantial burden for her. Notably, what explains why abortion is permissible according to this argument has

nothing to do with the moral status of the fetus or the mother, since Thomson *grants* that the fetus has a right to life. So, if Thomson's argument works, it would follow that whether it's permissible to kill some X cannot turn on X's moral status alone. Hence, (11) seems false.

Jeff's reply to this objection is that Thomson's argument falls short of establishing that it is permissible to kill the fetus in many cases of interest:

> What the Thomson argument justifies is the termination of the fetus's trespass against the body of the pregnant woman. Let us grant for the sake of argument that it can justify the killing of the fetus when that is necessary to terminate the trespass. But it does not follow that it can justify the killing of the fetus *via* abortion when that is *not* necessary to remove the fetus from the woman's body…. [T]he Thomson argument may show that it is permissible for the woman to remove the fetus by caesarian, even if that would be worse for it than a later, natural delivery; but it does not show that it is permissible to kill it *via* abortion when it is possible to have it removed alive, [esp. when this would not pose any significant threat to the mother].[2]

Since Jeff thinks most people would still believe that an abortion would be permissible in the cases where the Thomson argument does not establish its permissibility, he thinks the Consistency Argument will still go through for those cases.

6. Hassoun and Kriegel's Argument for the Permissibility of Infanticide

Since we otherwise wouldn't get to it, I thought I'd briefly state and criticize a much more theoretically loaded argument for the permissibility of infanticide that is found in the article by Hassoun and Kriegel. Simplified a little, they argue as follows:

I. It is impermissible to kill X only if X is a person.

II. X is a person only if X has the capacity for consciousness.

III. X has the capacity for consciousness only if X is capable of having mental states that are conscious.

IV. A mental state M of X's is not conscious unless X is aware of M.

V. X cannot be aware of a mental state M unless X is aware that *she herself* is in M.

VI. X cannot be aware that *she herself* is in M unless X has a concept of herself.

VII. There is a stage at which infants plausibly do not have concepts of themselves.

VIII. Therefore, there is a stage at which it is not impermissible to kill infants.

This argument seems to me to be unconvincing, because (V) is false for a familiar reason that Hassoun and Kriegel rather strikingly fail to anticipate.

Here is a simple argument against (V). Suppose I look at a speckled hen, and the side that is facing me happens to have 18 speckles. It is clearly true that I am aware *of* a hen whose

[2] McMahan (2007: 155 – 156).

facing side has 18 speckles. But, crucially, to say that I am aware *of some object X with a property P* is *not* to say that I am aware *that X has P*. I might have no concept of a hen, and I might not be able to tell exactly how many speckles the hen has without attending to each and counting them. To use a different example to avoid any semblance of begging the question against Hassoun and Kriegel, note that if I am looking at a bright dot in the night sky that happens to be Venus, it can be true to say of me that I am visually aware of Venus, even though it's false to say that I know that what I am looking at is Venus, and even false to say of me that I have the concept of Venus. So, generally speaking, *awareness-of* does not entail *awareness-that*. If this is true, then there is no reason to accept (V). (V) claims that the fact that I am aware of M entails that I am aware that M has some property (namely, the property of being a state that I am in). But this is false. Being aware of something doesn't require one to possess *any* concepts. Animals, I assume, are aware of things, but it would be implausible and controversial at best to assume that they all have concepts, since concepts require a level of cognitive sophistication that is beyond many animals that uncontroversially have conscious states.

1. Killing Non-Human Animals

We can organize our discussion of the ethics of killing non-human animals in much the same way as our discussion of the ethics of infanticide. For there seems to be a simple consistency argument that shows that killing some non-human animals should be impermissible if it is impermissible to kill all human beings (with the capacity for consciousness) or, alternatively, that killing some human beings should be permissible if it is permissible to kill all non-human animals. Here is the argument:

Consistency Argument II

1. There are some human beings (with the capacity for consciousness) that have all the same morally relevant intrinsic properties as certain non-human animals. (Compare, for instance, severely mentally disabled people, infants, or late-term fetuses with chimpanzees, dogs, and pigs.)

2. If two entities have all the same morally relevant intrinsic properties, it should be equally (im)permissible to kill them.

3. So, it should be just as (im)permissible to kill non-human animals as it is to kill those human beings (with the capacity for consciousness) that have the same morally relevant intrinsic properties.

The practical implications of Consistency Argument II turn on further assumptions. If we accept the following assumption, we have a case against killing some non-human animals:

4. It is impermissible to kill human beings that happen to have all the same morally relevant intrinsic properties as some non-human animals. (Examples: severely mentally disabled people and infants.)

(3) and (4) jointly entail that it is impermissible to kill some non-human animals – e.g., chimpanzees, dogs, and pigs. Alternatively, you might find Jeff's Consistency Argument for infanticide very compelling, and you might also believe that late-term abortions are permissible. If you did, and you therefore accepted that some cases of infanticide are permissible, you probably wouldn't find (4) as compelling, and wouldn't take Consistency Argument II to challenge our current widespread treatment of non-human animals.

How strong is Consistency Argument II? Well, it does knock down some proposals that might have seemed to justify our differential treatment of non-human animals and some human beings (e.g., the mentally disabled). After all, before thinking about Consistency Argument II, one might have been inclined to advance arguments with the following form:

The Quick Argument Schema

A. All human beings have fancy property F. (Perhaps F = rationality, or F = autonomy, or F = self-awareness, or F = linguistic ability, or F = cultural sophistication.)

B. No non-human animals have fancy property F.

C. Fancy property F is morally relevant.

D. So, F provides a basis for giving all human beings differential treatment from all non-human animals.

Reflecting on Consistency Argument II strongly casts doubt on the existence of any sound instances of the Quick Argument Schema. After all, infants and the mentally disabled make (A) false in many instances – e.g., in cases where F = rationality, autonomy, self-awareness, linguistic ability or cultural sophistication. Indeed, mentally disabled human beings also block appeals to the moral significance of the *potential* for having any of these properties. After all, some such beings do not seem in any relevant sense to have the potential for rationality, self-awareness, autonomy, linguistic ability or cultural sophistication. At the very least, they will not have these properties or the potential for them to a degree that would actually be sufficient to *also* make (B) come out true, since it is far from obvious that all non-human animals possess none of these properties to *any degree whatsoever*.

There is one further way of filling out the Quick Argument Schema on which (A) is true: just let F = *the genetic profile of Homo Sapiens*. The main problem with this suggestion is that it makes (C) false. One way of seeing this is to imagine that we find intelligent life forms on other planets whose psychological capabilities precisely mirror ours, and who have the same sorts of experiences of pain and pleasure. If we knew about the mental lives of such beings and could communicate with them, it just isn't plausible that we'd find it permissible to kill them except under very extreme circumstances mirroring those in which we'd find it permissible to kill people like you and me. A related problem with this suggestion is that it's hard to how genetic properties could be *intrinsically* morally significant. The obvious explanation of why certain genetic properties might be morally significant is that they, together with environmental factors, lead to certain phenotypic properties. But then we're back to square one, since, for every non-human animal, there will plausibly be some human being with phenotypic properties that have intuitively exactly similar moral significance.

Nevertheless, given that some might be willing to embrace the conclusion that it is not as wrong to kill infants or mentally disabled people as it is to kill normal adults, it may seem hard to get anything of huge practical significance out of Consistency Argument II alone.

But it may be easy to add some premises that are not quite as strong as (3) that would motivate changing our current practices towards non-human animals. For one thing, even if you don't think that it is as wrong to kill infants or mentally disabled people as it is to kill normal healthy adults, you probably don't also think that we have permission except in fairly extreme cases to go ahead and kill them. It would take a very strong competing reason to justify such killings – e.g., that killing them would enable us to save the lives of many normal healthy adults. So, if Consistency Argument II does establish that plenty of non-human animals are morally on a par with mentally disabled people or infants, it ought to follow that it would take very strong competing reasons to justify killing them.

Are there any such reasons? Perhaps there are, but they certainly don't seem to be ones that would justify our current carnivorous practices. As Jeff points out, the meat-eaters among us can't simply appeal to the pleasure they get out of eating meat as a justifying reason. A vegetarian diet could also afford considerable pleasure. They can at best appeal to the *difference* between this pleasure and the perhaps decreased pleasure of a vegetarian diet. This difference seems very minimal, and is arguably outweighed by the health benefits of a vegetarian or vegan diet. Could such a minimal difference really be enough to outweigh the interests that non-human animals have in continuing to live? More crucially, could such a minimal difference really be enough to outweigh the enormous suffering and pain to which non-human animals are currently subjected in factory farm settings? Barring a lot of further argument, the answer to the second question seems like a decisive 'no', and the answer to the first hardly seems to be a clear 'yes'.

What replies are available to the defender of current carnivorous practices? Three are particularly salient. The first is that abolishing the meat industry would threaten the livelihoods of hundreds of thousands of people, and massively disrupt the economy. So, even if there is no positive justification for ever *instating* practices that are analogous to current carnivorous ones, there is a strong reason against *stopping* these practices. As Jeff points out, this argument overgeneralizes in an embarrassing way: "[T]his argument…can be advanced on behalf of any large-scale social practice, however iniquitous. The same claim has been made in support of slavery in the South, the sale of tobacco products at home and abroad, and the sale of advanced weapons to tyrannous and illegitimate governments." And, as he notes, even if the argument didn't overgeneralize, it still wouldn't undercut a "creative program for the gradual elimination of the practice in which the government facilitates the process of economic conversion". So, the reply is doubly flawed.

Another salient reply concedes that there is a good argument for a society-wide abolition of current carnivorous practices, but denies that this yields a decisive reason for any particular individual like you or me to give up participation in such practices. After all, the meat-eater might reply: "If *I* gave up my carnivorism, it would not be sufficient to reduce the harm and suffering of animals to any noteworthy degree. So, *I* have no reason to abandon my contribution to the practice." One problem with this argument is that it overgeneralizes. An argument of the same form could be used to show that no one has any reason to vote, even if the voting system is perfectly in order and in support of majority rule. Indeed, an argument of the same form could be used to show that no individual member of a genocidal totalitarian state has a reason to make efforts to bring down the regime. This kind of argument cannot, then, be generally valid. Finally, as Jeff notes, even if one cannot alone significantly reduce the harm and suffering of animals by becoming a vegetarian, one can manage to "bear witness to the wrongness of the social practice of meat-eating, to set and example for others, and thereby to provide impetus and momentum to social action that *will* significantly diminish the harms that we as a society inflict upon animals."

A final reply would be to retreat to a more modest practice that Jeff calls *benign carnivorism*. Defenders of benign carnivorism would suggest that we should breed and rear animals in conditions in which they would be contented, and subject them to painless killings after they have lived reasonably long lives. The problem with this argument is that it isn't really a reply to the original objection. The point behind the original objection was that the minimal difference in pleasure that humans derive from eating meat cannot plausibly outweigh

animals' interests in continuing to live, at least insofar as continuing to live would continue to be a good thing for them. This objection remains unless it could be shown that conditions for animals could be improved so much beyond the conditions they would experience living free in the wild that it would outweigh their interests in living longer lives. This hasn't been shown. Although benign carnivorism may be less wrong than our current carnivorous practices, there is still no clear argument for its wholesale permissibility.

Notably, none of these arguments are meant to undercut the permissibility of painlessly killing animals for substantially greater causes than simply getting a tiny increase in pleasure. Painlessly killing animals for such purposes as xenotransplantation (i.e., the transplantation of organs from animals to humans) seems plausibly justifiable, since, as Jeff notes, "the harm that a person suffers in dying is normally considerably greater than that which an animal suffers [and so] the harm prevented through xenotransplantation would usually greatly outweigh the harm inflicted." But a crucial thing to notice about this point is that it has nothing special to do with non-human animals, since analogous points can be made in cases like *Altruistic Abortion* and *Healthy Newborn*, cases that Jeff takes to involve permissible killing. Indeed, Consistency Argument II could be used to show that we must accept that cases like *Altruistic Abortion* and *Healthy Newborn* (or perhaps analogous cases involving severely mentally handicapped people) are permissible if we accept that killing animals for the purposes of xenotransplantation is permissible.

2. Active vs. Passive (Voluntary) Euthanasia and Killing vs. Letting Die

We can start our discussion of euthanasia with a distinction between *active* and *passive* euthanasia. In active euthanasia, an individual is killed for the reason that this is all-things-considered better for him than allowing him to continue to live. In passive euthanasia, an individual is allowed to die for the reason that this is all-things-considered better for him than trying to preserve his life by various artificial means. Two questions to ask are:

Q1. Are (voluntary) active and passive euthanasia ever permissible?

Q2. Is (voluntary) active euthanasia at least as permissible as (voluntary) passive euthanasia?

There are strong cases that the answer to both of these questions is 'yes'.

One way of arguing for a positive answer to Q1 begins with the observation that it is pretty clear that suicide is morally permissible in cases where it is rational (since it would indeed be better for the suicide candidate to die rather than to live) and not worse for others. The argument then proceeds by showing that, at least in some cases, there is no principled reason why either passive or active euthanasia shouldn't be permissible if suicide would be permissible. To see this, let's consider two imaginable variations on a scenario that I borrow (with some modifications) from McMahan (2002).

Suppose that Bill is terminally ill and is suffering horribly. He has no dependents, relatives or friends who require his support, or who would oppose his committing suicide. And suppose that Bill rationally wants to commit suicide. Suppose, however, that Bill is disabled to such a degree that he cannot bring about his own death; all he can do by himself is just wait for his illness to take away his life, but this could last for months, and the agony would

only increase. Then, imagine a couple of variations on the case. In Variation I, Bill is on life support, and his death could be quickly brought about by terminating that life support. In Variation II, Bill is not on life support, but could be killed by a lethal injection. The question to ask is this. Would it be *morally worse* for a doctor who wants the best for Bill and heeds his competent request to die to terminate life support in Variation I or give a lethal injection in Variation II than it would be for Bill to do either of these by himself in some third scenario (call it Variation III) in which no disability prevented him from doing them?

It is very hard to see why we shouldn't say 'no' and regard the cases as morally equivalent. If so, we're forced to grant that some cases of passive euthanasia (as in Variation I) and active euthanasia (as in Variation II) are as permissible as cases of rational and impartially harmless suicide (as in Variation III). So, we seem to get clear positive answers to Q1 and Q2.

As Jeff says more elegantly about a real case involving Kevorkian, this time contrasting assisted suicide with active euthanasia:

> [Kevorkian's] earlier cases were instances of assisted suicide: he hooked people up to a device containing a lethal chemical, but the people themselves actually pressed the button that released the chemical into their bloodstream. By contrast, in a more recent case, the person…suffered from amyotrophic lateral sclerosis and was so disabled that it was difficult for him to push the button. He therefore asked Kevorkian to push it for him, which Kevorkian did…. *Most of us, on reflection, find it difficult to believe that it could make a momentous moral difference whether [the patient] pushed the button himself or whether Kevorkian pushed it for him. Yet that is all the difference between assisted suicide and killing amounts to.*[3]

In spite of all this, there is some pretheoretical inclination to give a negative answer to Q2. This might seem to be supported by the idea that killing is worse than letting die.

But, on more careful reflection, it seems like exactly the opposite should be true. Recall that we have seen three fairly plausible partial explanations of the wrongness of killing:

i. Killing a person is wrong in part because it deprives a person of all the good that would otherwise be in his future.

ii. Killing a person (with rationality and autonomy) is wrong in part because it seriously violates a requirement of respect.

iii. Killing a person is wrong because it violates his right to life.

(iii) cannot apply to the case of interest, since, in cases of voluntary (active) euthanasia, a person explicitly forfeits his right to life by competently requesting to be killed. (ii) cannot plausibly apply to the case of interest, since refusing to respond to someone's competent request to be killed and allowing him to endure much greater suffering and die a slower death by natural causes seems to show him *less respect* as a rational and autonomous being than granting his request. And (i) obviously cannot apply because, by definition, euthanasia only occurs when it would be better for a person to die than to continue living. So, none of

[3] McMahan (2002: 460).

the explanations of why killing is wrong turn into explanations of why actively euthanizing someone should be bad, or worse than allowing him to die.

Moreover, there just isn't any general motivation for thinking that killing should always be worse than letting die in the cases of interest to us. As Jeff points out, in cases where dying would be better for a person than continuing to live, the opposite should be true, since "benefiting a person is better than merely allowing him to be benefited", and hence "Kevorkian's actively bringing about [the] death [in the aforementioned case] was, if anything, better or more praiseworthy than merely allowing or enabling [the patient] to bring it about would have been".[4] So, if anything, appeals to more general distinction between *doing* and *allowing*, and to the fact that *doing good* is better than *allowing good*, should lead us to accept that killing is *better* than letting die in cases of voluntary euthanasia. So, the appeal to the difference between killing and letting die to give a negative answer to Q2 is misguided.

And, at the end of the day, it isn't clear if it is true that killing is worse than letting die *even when the death would be bad*. James Rachels famously argued for the irrelevance of the general distinction by noting our intuitions about the following pair of cases:

> *Smith's Case*. Smith stands to gain a large inheritance if anything should happen to his six-year-old cousin. One evening while the child is taking his bath, Smith sneaks into the bathroom and drowns the child, and then arranges things so that it will look like an accident. No one is the wiser, and Smith gets his inheritance.
>
> *Jones's Case*. Jones also stands to gain if anything should happen to his six-year-old cousin. Like Smith, Jones sneaks in planning to drown the child in his bath. However, just as he enters the bathroom, he sees the child slip, hit his head, and fall face-down in the water. Jones is delighted; he stands by, ready to push the child's head back under if necessary, but it is not necessary. With only a little thrashing about, the child drowns all by himself, 'accidentally', as Jones watches and does nothing. No one is the wiser, and Jones gets his inheritance.[5]

As Rachels notes, if we really think that the distinction between killing and letting die has fully general significance, we ought to regard *Jones's Case* as less morally reprehensible than *Smith's Case*. But it seems we don't. So, he reasons, it seems there is no reason to accept that the distinction between killing and letting die always tracks a morally significant distinction.

One might be left wondering why it seemed like there was a general morally significant distinction here to begin with. My own suspicion is that particular cases in which killing is worse than letting die are just cases in which explanations (i), (ii) or (iii) of the wrongness of killing come into play. If someone didn't ask for us to kill him but he is going to die anyway, we do wrong by killing him rather than simply allowing him to die because we violate the respect demanded by his autonomy, or violate his right to live as long as he wants to live. And if someone would still get some enjoyment out of life but is going to die soon, killing him right away is obviously worse than letting him die because we're taking away that last bit of enjoyment (as well as violating the respect violated by his autonomy and his right to live as long as he wants to live). So, explanations (i), (ii) and (iii) seem sufficient to explain

---

[4] McMahan (2002: 461).

[5] These cases are taken *verbatim* from Rachels (1986: 112).

intuitions in which killing is clearly worse than letting die. But this is compatible with the insignificance of this distinction in cases where explanations (i), (ii) and (ii) *do not apply* – e.g., in comparisons between voluntary passive euthanasia and voluntary active euthanasia.

It is also worth noting that it is easy to conflate questions of *impermissibility* with questions of *blameworthiness* and, relatedly, assessments of *acts* with assessments of *agents*, even though there is a distinction between the two. Someone can do something all-things-considered wrong (and hence impermissible) but be blameless for it, and can do something all-things-considered right (and hence permissible) but not be praiseworthy for it. For instance, if Jones had misleading evidence that giving a certain drug to someone would cure his ills when it would in fact kill him, Jones wouldn't be blameworthy for giving the drug to the person in the belief that he would be helping him, even though there is, from an impersonal and fully informed point of view, nothing to recommend about the act that he performed: after all, he killed the person. Alternatively, if Jones had misleading evidence that giving a certain drug to someone would kill him when it would in fact cure all his ills, and Jones gave the person the drug *because* he wanted to kill him, Jones is clearly not praiseworthy for his act, even though there is, from an impersonal and fully informed point of view, something to recommend about the act that he performed: after all, he saved the person.

I suspect that part of the reason why it is tempting to find cases of killing worse than cases of letting die is that it is easy to conflate blameworthiness with impermissibility. Someone who had influence over whether some bad upshot would occur and who exercised his influence by intentionally causing that upshot to occur is obviously more blameworthy than someone who did not want the upshot to occur but who could have prevented it from occurring. We are, after all, justified in assessing agents by their intentions, and there is a difference here. But our assessments of acts and omissions need not be colored by our assessments of the people who have causal influence over them. I'd hypothesize, then, that we can *explain away* the feeling that some examples of killing are worse than some examples of letting die by conceding that there is often a basis for regarding killers as more blameworthy than those who simply let death occur when it could have been prevented while insisting that the acts/omissions in question are equally impermissible.

This prediction is confirmed by our intuitions about the two cases that Rachels considers. What is unusual about his cases is that both Smith and Jones are terrible agents with blameworthy intentions; usually, an agent in a paradigm case of killing will have more blameworthy intentions than an agent in a paradigm case of letting die. Rachels's cases helpfully prevent us from conflating impermissibility and blameworthiness because there is no difference in blameworthiness. And, as predicted, if we fix this confounding factor, it seems we aren't inclined to think that there should be any further difference in permissibility.

3. A Small Wrinkle: Passive vs. Active <u>Involuntary</u> Euthanasia

We do have to be a little careful in generalizing the idea that there is no significant moral distinction between passive and active euthanasia. So far we've only been considering cases of *voluntary* euthanasia – i.e., cases in which dying would not only be better for someone, but in which they want to die and competently request to be killed or allowed to die. Things get more intricate when we consider cases of *involuntary* euthanasia – i.e., cases in which dying

would be better for someone, but in which they do not want to die. For few people would find the following argument persuasive:

I. If dying would be better for X, then even if X does not want to die, it would not be morally wrong to allow him to die and hence to let him be passively euthanized (say, by failing to provide him with treatment that might allow him to live a bit longer, but where this life would be full of egregious suffering and would not be worth living).

II. There is no distinction in permissibility between passive and active euthanasia.

III. Therefore, if dying would be better for X, then even if X does not want to die, it would not be morally wrong to kill him and hence actively euthanize him.

Many people would not be willing to accept (III); as I explained last week, I think that it is hard to see why (III) really should be rejected, but I'll set that concern aside for the moment. Nevertheless, (I) still seems compelling, or at least much more compelling than (III). So, to prevent the derivation of (III), we may have to qualify the conclusion from the last section:

II*. There is no moral distinction in permissibility between voluntary passive and voluntary active euthanasia.

(III) does not follow from (I) and (II*).

Besides simply wanting to avoid the derivation of a conclusion that many people would not be willing to accept, why should we restrict our conclusion from the last section in the way suggested by (II*)? Well, suppose that you do reject (III). Why might you be inclined to reject (III)? The only clear reason that I can see is that you'd think that it violates X's rights: even if X's life isn't worth living, X has a right to life, and this right must be respected unless X forfeits it. Crucially, this reason for rejecting (III) does not turn into a reason for rejecting (I). For, as we saw when we discussed Thomson, the following argument is invalid:

a. X has a right to life.

b. Therefore, X has a right to be given whatever would keep him alive by anyone.

Even if you agree that the (a – b) argument is invalid, you could consistently cling to:

a. X has a right to life.

b*. Therefore, X has a right not to be killed against his will.

If you did reject (a – b) but clung to (a – b*), you would have a reason for thinking that (III) is false that isn't also a reason for thinking that (I) is false. So, there are principled reasons for restricting our conclusion from the last section to (II*), since there are principled reasons that distinguish between the permissibility of involuntary active euthanasia and involuntary passive euthanasia. Basically, these reasons have to do with exactly what further rights the right to life implies: perhaps the right to life implies the right not to be killed unjustly without implying the right to be given whatever will preserve one's life.

1. A Recap of the Taxonomy of Normative Theories from Last Time

Near the end of the last section meeting, someone asked me to explain how preference utilitarianism differs from just-plain-old utilitarianism. I replied by quickly sketching a general taxonomy of consequentialist theories, indicating where preference utilitarianism falls in the hierarchy, distinguishing some subspecies of it, and illustrating some versions of utilitarianism that differ from it. Although this is a class in *applied* ethics and not a class on higher-level ethical *theory*, I think it is still appropriate to make everyone in a less theoretical class familiar with some of the basics of the higher-level theoretical terrain. So, I thought I'd briefly put some of the stuff that I was saying very quickly last time on this handout so that people can refer to it later and get used to the precise ways in which terms like 'consequentialism' and 'utilitarianism' are to be used.

Broadly speaking, general normative ethical theories fall into two categories: *consequentialist* theories and *deontological* theories. Since deontological theories are typically defined in opposition to consequentialist theories, it is best to start with a taxonomy of consequentialist theories.

There are strikingly many versions of consequentialism. One general distinction is between *act* consequentialism and *rule* consequentialism. An act consequentialist theory of rightness will claim that the rightness of some act performed in some circumstances C turns entirely on the consequences of that very act in C. In contrast, a rule consequentialist theory of rightness will claim that the rightness of some act performed in some circumstances C turns entirely on the consequences of general acceptance of the policy of allowing this kind of act to be performed in circumstances like C. These two theories obviously have different implications. Perhaps the objective probability that some act A would have good consequences in C-type circumstances is exceedingly low. Nevertheless, suppose that someone performs A, and by chance A happens to have really great consequences. An act consequentialist might claim that this act is right, while a rule consequentialist might claim that this act is wrong, since the general policy of allowing A in circumstances like C isn't actually so great, given the objective chances.

Another general distinction is between what are called *actualist* consequentialist theories and *expectabilist* consequentialist theories. A strong version of actualist act consequentialism would claim that it is right for some person P to perform some act A if and only if (this is abbreviated as "iff") P's A-ing *in fact* brings about the best consequences. A strong version of expectabilist act consequentialism would claim that it is right for some person P to perform some act A iff P *expects* that A-ing would bring about the best consequences. These two theories make different predictions, since someone's expectations could be mistaken.

I called both of these theories 'strong' versions of consequentialism. Why? Well, here are two different versions of actualist act consequentialism, the first of which is much stronger:

> *Maximizing Actualist Act Consequentialism*: It is right for P to A iff P's A-ing in fact brings about the *best* consequences.

> *Satisficing Actualist Act Consequentialism*: It is right for P to A iff P's A-ing in fact brings about consequences that are *good enough*.

People often retreat from maximizing to satisficing versions of consequentialism because maximizing consequentialism seems to be an extremely *demanding* theory. After all, it is a logical consequence of maximizing actualist consequentialism that if P performs an act that brings about consequences that are just ever so slightly less good than some optimal alternative, he acts *wrongly*. Suppose, for instance, that we could numerically measure goodness, and that the optimal act brings about 1,000,000 units of goodness, while the act that P actually performs brings about 999,999 units of goodness. The theory in question entails that P acts wrongly. If this theory were true, it is highly likely that most of us act wrongly all the time. This might look counterintuitive. So, some people suggest that we retreat to satisficing consequentialism. Of course, the big problem with satisficing consequentialism is that there is a burden on its defenders to explain just how much production of good consequences is good enough. And there is a significant worry that there is no nonarbitrary answer to this question.

Let's set that aside, and focus on maximizing actualist consequentialism for simplicity's sake. There are many different varieties of maximizing actualist consequentialism. The main factor that distinguishes these theories is the underlying theory of optimality on which they rest. Utilitarians are consequentialists who endorse the following theory of optimality:

> *Utilitarian Account of Optimality*: Some consequences are the best iff, were these consequences to be actualized, well-being would be maximized.

What is well-being? That's a question on which utilitarians disagree quite a bit. There are lots of subspecies of the Utilitarian Account of Optimality. For simplicity, we can focus on:

> *Hedonism*: Well-being consists in a high pleasure-to-pain ratio.
>
> *Preferentialism*: Well-being consists in a high preferences-satisfied-to-preferences-frustrated ratio.
>
> *Pluralism*: Well-being consists in possessing a high ratio of intrinsic goods (friendship, knowledge, physical health, pleasure, etc.) to intrinsic bads (strife, ignorance, physical unfitness, pain) in one's life.

So, the long answer to the question I was asked last week is that just-plain-old utilitarianism is neutral on whether Hedonism, Preferentialism or Pluralism is true, while preference utilitarianism explicitly endorses Preferentialism.

There are also non-utilitarian versions of consequentialism, but I'll set them aside and turn to a brief discussion of deontological theories. As I said, deontological theories are usually defined in contrast to consequentialist theories. Most deontologists would, for instance, accept the following clearly non-consequentialist claim:

> *Deontological Dictum* (DD). For at least some acts A and circumstances C, it can be right to perform A in C even if some alternative to A has substantially better (actual

> or expected) consequences in C, and even if the general policy of A-ing in C-like circumstances would, if adopted at large, would lead to substantially better (actual or expected) consequences.

Deontological theories differ on the score of exactly how many types of acts instantiate DD, and on the score of what explains why these acts instantiate DD.
A good way of bringing out the contrast between theories that accept DD and theories that reject DD is by considering the following kind of case:

> *Fat Man on the Bridge.* A trolley car is speeding down the tracks and is just about to pass under a bridge. Farther down the tracks, seven people have been tied down by a maniac and will be killed by the trolley if you, who are standing on top of the bridge, don't do something. The only way you could stop the trolley is by hurling something massive in front of it. You're too slender and couldn't do anything by jumping in front of it, and there aren't any big boulders or anything else inanimate that you could throw in front of it either. There is, however, an enormous man standing directly above where the train will pass under. You don't have enough time to persuade him to jump, but you could run and push him off the bridge. It is certain that, if you did this, you would stop the trolley and thereby save the seven people further down the tracks.

Pushing the fat man off the bridge would seem to have substantially better consequences than failing to do so in this case, as would the general rule that this act instantiates (i.e., *kill one if it would save seven*). So, at least on many versions of act consequentialism, it will follow that it would be right to push the fat man off the bridge. This is a result that many find intolerable, and provides a seemingly clear argument for the Deontological Dictum. What explains why we're willing to reject consequentialist reasoning here? Deontologists point to:

> *Mere Means Principle.* It is never permissible to use another person as a *mere means* to some end.

Deontologists often endorse this kind of principle, and claim that it explains our intuitions about *Fat Man on the Bridge*.

There's obviously a lot more that can be said about consequentialism, deontology, and the disputes between the two theories, but I'll leave it at this for the purposes of this class.

2. The Cons of Legalized Euthanasia

Even if one accepts that active and passive voluntary euthanasia are morally permissible, one needn't automatically accept that it would be a good idea to legalize them, at least without some considerable restrictions. There are many reasons why one might be inclined to draw a boundary line here, but most of them fall into one of the following categories, each of which I'll discuss at greater length:

> 1. *The Slippery Slope.* Some worry that if we legalize clearly morally acceptable forms of euthanasia, we will set foot on a slippery slope that will take us to legalizing morally unacceptable practices like involuntary euthanasia. There will, perhaps, be no barrier that will prevent us from eventually legalizing involuntary euthanasia, or from mistakenly regarding as "euthanasia" what may really be worse for the patient

and legally allowing him to be killed when it wouldn't really be a good case of mercy-killing.

2. *Possibility of Abuse.* Some worry that any law that is robust enough to allow all morally permissible will be gappy enough to permit some morally unacceptable cases. In other words, there might be loopholes in any law that would be strong enough to permit the good cases of euthanasia that could be abused.

3. *Possibility of Mistakes.* Some worry that if we legalize morally permissible types of euthanasia, we might make hasty and mistaken judgments about whether individual cases exemplify these types. There might be cases of misdiagnosis, cases in which someone might have unexpectedly recovered, or cases where a cure might be discovered right after the person is killed. We wouldn't want any of these things to happen, and some people think this is enough to constitute a case against the legalization of euthanasia even under highly restricted conditions (e.g., in which the patient is believed to have a terminal illness and to be incurably suffering to a degree that truly makes life intolerable).

4. *Bad Side-Effects: Pressure.* Some worry that legalizing clearly morally permissible forms of euthanasia will have adverse side-effects. Perhaps people will end up feeling pressured into giving consent to be euthanized when, deep down, they would prefer not to be euthanized. If we ended up accepting some list of conditions under which a life would be regarded as no longer worth living, a patient who constitutes a financial drain or an emotional burden on his family may feel pressured to take himself off the scene even though he doesn't really agree that his life is no longer worth living.

After talking a bit more about each of these concerns, I'll address the hard question of what the legal restrictions on permissible voluntary active and passive euthanasia should be, and whether there should be a legally recognized set of conditions that must be satisfied for a life to be regarded as no longer worth living. I'll also talk a bit about an interesting proposal that James Rachels makes about how euthanasia should be legalized.

2.1. *Arguments from the Slippery Slope*

The slippery slope argument is perhaps the most common one raised against the legalization of euthanasia. There are two forms that this argument sometimes takes among its defenders:

- *Psychological Slippery Slope.* If we legalize morally permissible practices like voluntary passive and active euthanasia, we will let down our inhibitions against involuntary euthanasia and killing people whose lives we may mistakenly think aren't worth living, and will inevitably end up legalizing these bad practices, too. Since we shouldn't legalize these bad practices, we shouldn't legalize any kind of euthanasia at all.

- *Logical Slippery Slope.* If we legalize morally permissible practices like voluntary passive and active euthanasia, there will be no principled logical reasons not to legalize involuntary euthanasia and the killing of people whose lives we may mistakenly think aren't worth living. So, if we want to have a logically consistent law, we couldn't legalize the good practices without legalizing the bad ones. Since we shouldn't legalize the bad ones, we shouldn't legalize euthanasia at all.

The second version of the argument is clearly mistaken, since there are clear and, indeed, obvious reasons why involuntary euthanasia and the killing of people whose lives really *are* worth living are impermissible that do not apply to voluntary active and passive euthanasia. The first version of the argument is less clearly mistaken, since it relies on a psychological hypothesis that isn't obviously disconfirmed.

Some people think that this hypothesis is clearly confirmed, but for bad reasons. One case that people routinely bring up is that the Nazis' program of mass killing had its origins in something that they called 'euthanasia': they thought there was such a thing as a life not worth living, and the people to whom this thought was initially applied were not members of an ethnic group, but simply the severely and chronically sick. Here there was in fact a clear slippery slope down which the Nazis careened. Is this compelling evidence for the hypothesis on which the psychological slippery slope argument relies?

It isn't, for reasons that James Rachels usefully explains in his book:

> Are we to believe that Hitler and his followers were at first an ordinary group of people who permitted mercy-killing from a sense of compassion? And that this led them, in less than a decade, to be transformed into the monsters of concentration camps? Of course this is not what happened…. Among the Nazis, there was never any thought of killing as a compassionate act for the benefit of suffering terminal patients; indeed, this was not even used as a false excuse when they would lie about what they were doing…[n]or was there ever any thought of securing the permission of the victims. The sterilizations as well as the killings were completely involuntary. Where, then, is the analogy with the real euthanasia movement?[6]

To have a real empirical case for the psychological assumption made by the slippery slope argument, we would need an example where voluntary passive and active euthanasia were legalized out of a sense of compassion, and where this eventually led to legally accepted clear cases of involuntary euthanasia or killings that weren't really better for the patients. This hasn't been shown. It isn't, for instance, illustrated by the case of the Netherlands, since the farthest they have gotten from voluntary active and passive euthanasia are some cases of non-voluntary euthanasia – cases in which patients were unable to express their preferences, and had expressed no prior preferences about how to be treated in these conditions. And there are perfectly good reasons for thinking that these types of cases can be morally permissible.

So, the psychological argument rests on an assumption for which there is no real evidence.

2.2. *Arguments from the Possibility of Abuse*

Arguments against the legalization of voluntary active and passive euthanasia that proceed from concerns about abuse take many forms. One argument is what we might call the *Argument from Bad but Unknowable Motivations*:

> 1. If voluntary active euthanasia were legalized, there would inevitably be legally permitted cases in which the person who provides the (putative) euthanasia has very corrupt reasons for it – say, he simply wants to get his inheritance and

[6] Rachels (1986: 177 – 178).

doesn't give a damn for the well-being of the person asking for death – though he very effectively hides those reasons.

2. There shouldn't be legally permitted cases in which the person who provides the (putative) euthanasia has motivations that are this corrupt.

3. Therefore, voluntary active euthanasia shouldn't be legalized.

Premise (1) may very well be true. It is hard to see how the law could effectively detect cases in which the (putative) euthanizer secretly only grants the wish of the (putatively) euthanized person for reprehensible reasons, but in which he is careful to prevent any evidence of his corrupt motives from showing up. It is less obvious, however, that the inference from (1) and (2) to (3), or, for that matter, premise (2) is acceptable. I suspect, for instance, that this is a completely general problem: *any* legal setup will contain loopholes that permit someone to perform an act that is perfectly permissible with intentions that are clearly reprehensible. Moreover, given the distinction we've been making between evaluations of acts (permissibility/impermissibility) and evaluations of agents (praiseworthiness/blameworthiness), and the fact that what's really at stake here is the potential legal permission of morally impermissible acts, it's not clear that there is really a serious worry behind the argument. People's motives might be inscrutable to us; what matters is that they never perform an act that is itself impermissible. The Argument from Bad but Unknowable Motivations does nothing to establish that such acts will be allowed.[7]

Another worry about abuse is presented by the *Argument from Badly Persuaded Consent*:

4. If voluntary active euthanasia were legalized, there would inevitably be legally permitted cases in which someone is euthanized who came to desire and request death only as a result of illegitimate persuasion, brainwashing or the like.

5. There shouldn't be legally permitted cases in which someone is euthanized who came to desire and request death only for these reasons.

6. Therefore, voluntary active euthanasia shouldn't be legalized.

This argument is a little more compelling than the last, since here a case could be made that the law may end up permitting some genuinely impermissible acts. Euthanizing someone

[7] Admittedly, making use of the act/agent distinction is trickier when we turn from morality to criminal law, since it's arguable that punishment should turn more on whether someone was blameworthy than on whether someone brought about some bad upshot. In some cases of attempted murder, for instance, nothing bad actually happens. The person whom the murderer was trying to murder may not in any way be harmed, and may not even realize that there was anybody trying to harm him. Perhaps, for instance, the murderer was a sharpshooter who was caught by a watchman in a tower just as he was about to fire his gun at some people on the street below. Here nothing bad actually happens, and so, at least on *some* normative theories, no impermissible act has been performed. This would follow, for instance, on a consequentialist theory on which the wrongness of an act turns on whether it had any bad actual consequences. But not everyone will agree that the sharpshooter ought to get away scot free in the imagined case! If so, then we can't assume that the criminal law ought to address only impermissible acts; it also ought to address some *merely blameworthy* acts. So, perhaps the defender of the Argument from Unknowable Bad Motivations would complain about what I've said. Still, I think the point about the implausible overgeneralizability of the argument holds: this is a broader problem, and to avoid it would almost certainly require a law that was prohibitive to an extreme and unreasonable degree.

who only wants to be euthanized because he's been brainwashed would not seem to show respect for his autonomy, particularly if the people doing the euthanizing are the very same people who did the brainwashing. It is far from obvious that this would not be morally wrong. Here, too, though, I suspect that there may be another completely general problem that could only be fully avoided by an impossibly prohibitive law. A clever ex-rapist might discover ways of brainwashing potential victims into having consensual sex with him, and may be able to prevent any evidence from accumulating that would be sufficient to establish that he was clearly violating his victims' autonomy. One might think that this is wrong, but there seems to be little we can do to design a law that would *absolutely guarantee* that this would never happen. A law that absolutely guarantees prevention of all wrongdoers from going scot free for their wrongdoings may not be feasible. The best one can do is weigh the costs of having a fairly prohibitive law against the benefits of having a more liberal one. And nothing has been shown that the benefits of allowing voluntary active euthanasia, at least in a sufficiently careful fashion, would be outweighed by the costs.

2.3. *Arguments from the Possibility of Mistakes*

The next major argument against the legalization of voluntary active and passive euthanasia worth taking seriously is the argument from the possibility of mistakes. This argument can be stated in the following form:

> 7. If voluntary active and passive euthanasia were legalized, there might be legally permitted cases in which someone is killed or allowed to die on the mistaken belief that s/he would not recover from her terminal illness, or (more plausibly in the case of a non-terminal but unbearably painful medical problem) that no medical treatment would be come available during the lifetime s/he would otherwise have that would be sufficient to make his or her life worth living again.
>
> 8. We should make certain that no such cases ever come about.
>
> 9. The only way to do this is to refuse to legalize any kind of euthanasia.
>
> 10. So, we should refuse to legalize any kind of euthanasia.

The first thing to realize in reflecting on an argument like this is that failing to legalize any kind of euthanasia is itself a choice that is very costly, since it would guarantee that many people will not be able to exit a life that is unquestionably not something that they would want to continue living, and that would unquestionably be worse and, indeed, intolerable for them to continue living. So, a question immediately arises about how this big cost compares with the cost of the possible mistakes to which this argument is gesturing. There are many things we can do to vastly decrease the objective probability that these mistakes will occur, and hence to vastly decrease the costs that would be engendered by legalizing voluntary euthanasia. If we really did these things, the real costs of potential error will arguably less than the costs of failing to legalize any kind of euthanasia. So, there seems to be a good reason to reject the assumption in this argument that we should make certain that no potential errors are ever legally permitted: if it is more costly to do this than to allow some very unlikely potential errors and at the same time relieve a huge amount of suffering, the assumption is simply false.

The next thing to realize in reflecting on this argument is that it's very hard to see why it doesn't overgeneralize in a clearly implausible way. To see this, consider an argument that has exactly the same form that purports to show that it shouldn't be legally permissible for people to refuse medical treatment:

> 7*. If the refusal of medical treatment by patients were legalized, there might be legally permitted cases in which someone is allowed to refuse treatment who will definitely have a much worse life without it.
>
> 8*. We should make certain that no such cases ever come about.
>
> 9*. The only way to do this is to not legalize the refusal of medical treatment.
>
> 10*. So, we should not legalize the refusal of medical treatment.

This argument is clearly absurd. But the problem is that the reason why this argument is absurd provides a direct objection to the argument against legalized active and passive euthanasia. This argument fails because we think that if we do not allow patients to refuse medical treatment in cases where this would definitely not be good for them, we fail to respect their autonomy. This is simply a balance of costs: we have a choice between failing to respect people's autonomy and failing to forcibly make their lives better, and it seems like we are ultimately willing to regard the second as a less serious cost than the first. But we have, after all, only been considering arguments for *voluntary* euthanasia, and voluntary euthanasia involves a persistent request and desire on behalf of the patient to be killed or to be allowed to die. A refusal to take such a request and desire seriously is just as much a failure to respect a patient's autonomy as a refusal to allow a patient to decline medical treatment. So, it seems like the reasons that block the absurd argument against the legalized refusal of medical treatment also block the argument against legalized voluntary euthanasia that we've been considering.

If a person decides to accept death in order to avoid a future life he reasonably believes will be unendurable, he should be aware that there is a tiny probability that the prognosis he has been given is mistaken, or that a cure will be found immediately after he dies. If he is aware of these possibilities and decides to die nonetheless, we cannot claim that his decision is irrational. It should be stressed that this is comparable to decisions that the rest of us make all the time. One may decide to undergo surgery for a nonlethal ulcer condition knowing that there is a small risk of death in subjecting oneself to total anesthesia. One risks losing the possibility of a good future life in order to avoid the pain of the ulcer. This is similar to a choice to accept the loss of a tiny probability of a good future (through the possibility of cure, etc.) through euthanasia in order to avoid the pain of whatever condition it is that makes one's life unendurable. If we don't think the possibility of mistakes in these other improbably risky practices is sufficient to make us refuse to legally allow them, we shouldn't think any differently about voluntary euthanasia.

2.4. *Bad Side Effects: Pressure*

Another major concern about the legalization of voluntary euthanasia is that its availability may end up pressuring people into giving consent to be euthanized when, if the practice

weren't so commonly accepted and, indeed, expected of people in certain conditions, they would really prefer not to be euthanized.

This concern is particularly vivid if we end up accepting some specific list of substantive conditions under which a life would be regarded as sufficiently no longer worth living to make voluntary euthanasia permissible. A patient in these conditions who also constitutes a financial or emotional burden on his family may feel that he ought to take himself off the scene. On the other hand, if we don't accept a list of specific substantive conditions under which a life would be regarded as sufficiently no longer worth living to make voluntary euthanasia permissible, we seem to open up the possibility of killings that are in fact against the interests of the persons killed. Moreover, there is a practical burden on physicians if such a list of conditions aren't accepted, since they will have less assurance that they are not being asked to kill someone when to do some would actually harm that person.

So, there's a dilemma: either (i) we accept a specific list of conditions under which a life would be regarded as sufficiently no longer worth living to make voluntary euthanasia permissible, and there will be a problematic kind of implicit pressure on potential candidates for euthanasia, or (ii) we do not accept such a list of conditions and thereby increase the risk of enabling voluntary killings that are not in fact in the best interest of the persons killed, as well as decrease the personal assurance that providers of euthanasia would have that they are doing the right thing.

The main thing to note is that exactly the same point can be made here that was made about the possibility of mistakes. While, if we take either horn of the dilemma, there may be costs, we have to ask whether these costs would be at least as great as the costs of failing to have any legalized practice of euthanasia at all. This has not been established, and it seems very doubtful that it could be established. This is particularly clear in the case of the first horn of the dilemma, since, by assumption, being killed would genuinely be a case of euthanasia and hence be better for the person in that case. Really the only serious problem on the first horn of the dilemma is that the person's consent may not be a reflection of his deepest desires, so that we would be failing to respect his autonomy in this case. But, since the practice here would still be voluntary euthanasia, it's not as though there are going to be cases where the pressure amounts to any sort of coercion: the patient will always have the choice to refuse.

And this brings out another way in which the reply to this concern can mirror the reply to the argument from the possibility of mistakes. Obviously, when a doctor strongly recommends treatment to a patient, his suggestion will typically put a bit of pressure on the patient to heed the recommendation and accept the treatment. Perhaps, if the doctor hadn't explicitly recommended a certain kind of treatment of which a patient was already aware and that the patient could have sought out by himself, the patient will be pressured into having a desire that he might not otherwise have had: perhaps, if he hadn't received the recommendation, he would have never sought out the treatment by himself, even if he was aware that it was very effective, and knew all the reasons why a doctor would recommend it. We do not conclude from this fact that doctors infringe upon patients' autonomy when they strongly recommend treatment, and conclude that they ought never to make the recommendation at all. If this kind of pressure isn't really a very serious problem, it's hard to see how the kind of pressure posed by the first horn of the dilemma we're considering is a

serious problem since, by assumption, we're talking about genuine euthanasia here – cases in which it really would be better for a patient to be killed.

So, in short, it seems like there isn't a sufficiently big worry presented by the dilemma to make the alternative of refusing to legalize euthanasia at all more attractive than, say, the first horn. If so, the version of the argument from bad side-effects we're considering fails.

2.5. *What Should the Restrictions Be? Should We Have a Set of Conditions for Quality of Life?*
Even if the arguments against the legalization of voluntary euthanasia fail, they do leave us with an important question that it may be extremely difficult to resolve in any satisfactory way. To avoid worries about abuse and the psychological slippery slope, we arguably do want to impose *some* restrictions on the conditions under which voluntary euthanasia should be legally permitted. What should these restrictions be? Should we have some set list of substantive conditions under which we could legally declare that a life is no longer sufficiently worth living that voluntary euthanasia could be permitted? What would those conditions be, and how would we decide?

One restriction that one often sees showing up in the lists of advocates of legalized euthanasia is that the patient be suffering from a terminal illness. This condition seems much too restrictive, since euthanasia is arguably much more desirable in cases where there would otherwise be "no end in sight". An example that Jeff discussed much earlier in the class in a different context is the case of a locked-in patient who can do no more than blink or tap a finger to communicate, and who is suffering from unbearable, unremitting and incurable pain, but who has no terminal illness. If anything looks like a life no longer worth living, this would be it. So, the terminal illness condition seems to be much too strong as a necessary condition for the legal permissibility of acquiring voluntary euthanasia.

A less restrictive necessary condition might be that the patient must be suffering from an *incurable* illness whose characteristic effects (e.g., incredible agony and suffering) make that person's life not worth living. Even this condition seems too strong, at least if 'incurable' is being used in anything like its intuitive, ordinary sense. Some forms of cancer are not strictly incurable: there are treatments that could in principle cure them. What matters is the probability and the risks involved in the treatment: the probability might be very low, and the treatment could vastly diminish the quality of the patient's life; just think of what's involved in certain varieties of chemotherapy. It certainly seems like we would want euthanasia to at least be an option for patients in these circumstances, but this would be ruled out by the condition at issue.

We could try to retreat to the claim that the condition must simply be *practically* incurable in the sense that all known potential cures have a very low success rate and involve substantial risk. But it's not clear how much this helps if our goal was to try to get a list of *usable legal guidelines to follow*. The law would have to settle on some sufficiently low success rate or sufficiently high risk, and it's not clear how we could choose one that would be sufficiently precisely defined to be both useful and that would not be overly restrictive.

Should it be a necessary condition that the patient be in unremitting and incurable pain? This is dubious if we go back to one variation on the case that Jeff discussed much earlier in the semester. Being a locked-in patient literally can do nothing more than blink would, to

many people, be intolerably horrible even if it didn't involve pain. To tell a locked-in patient who would indeed find this state an intolerable one to be in for an indefinite number of years that he is not permitted to choose to be euthanized would seem to be much too restrictive.

Once, however, we've ruled out unremitting and incurable pain, incurable illness, and terminal illness as necessary conditions, it's very hard to see how we could find any other neat necessary conditions that wouldn't simply be highly subjective. Arguably the crucial factor that makes it voluntary euthanasia seem permissible in the case of the locked-in patient who isn't in pain is simply that the patient himself finds his state intolerable. After all, as Jeff also pointed out in discussing this case earlier in the semester, some locked-in patients might actually enjoy their state: maybe they can still do math problems in their head, and this is the most enjoyable thing that they can imagine. So, it's not as though the locked-in state *by itself* is sufficient to make life not worth living: some remotely conceivable locked-in patients would like their lives.

And once this is appreciated, the prospects for getting any neat necessary condition for the legal permissibility of euthanasia seem to look rather dim. Perhaps the best that we can do is simply have some long, cumbersome list of sufficient conditions for legal permissibility, none of which are necessary. But this raises a concern about how effectively we could ever deploy such a list in practice, and make the decisions that need to be made.

2.6. *Rachels's Proposal*

There is, then, a concern about whether the restriction needed to avoid all the biggest worries about the consequences of legalizing euthanasia would be so cumbersome as to simply defeat the point of allowing euthanasia. As Rachels puts it:

> The problem with these proposals is, obviously, that they are so elaborate, and take so much time, that they are hardly conducive to the 'quick and easy death' that is the whole point of euthanasia…. As the American lawyer Yale Kamisar remarks, 'the legal machinery is so drawn-out, so complex, so formal and so tedious as to offer the patient far too little solace'.[8]

So, we're confronted with a dilemma. Either we impose lots of restrictions and eliminate the whole point of allowing euthanasia, or we get rid of the restrictions and get hit with the worries like abuse and the psychological slippery slope. At the end of his book, Rachels proposes a way of avoiding the first horn of the dilemma, but it's unclear to me whether it also manages to avoid the second. Here's his proposal in his words:

> [M]y suggestion for legalizing euthanasia is that a plea of mercy-killing be acceptable as a defense against a charge of homicide in much the same way that a plea of self-defense is acceptable. When people plead self-defense, it is up to them to offer evidence that their own lives were threatened and that the only way of fending off the threat was by killing the attacker first. Under my proposal, someone charged with homicide, in any of the varieties this charge may take, could plead mercy-killing; and then, if it could be shown that the victim while competent requested

[8] Rachels (1986: 183).

> death, and that the victim was suffering from a painful terminal illness, the defendant would also be acquitted.[9]

This avoids complications on the end of the person who requests euthanasia. But does it really avoid the worries posed by arguments like the argument from abuse or the psychological slipper slope argument? This seems unclear. To bring this out, consider an addendum Rachels makes:

> If this proposal were adopted, it would *not* mean that every time euthanasia was performed a court trial would follow. In clear cases of self-defense, prosecutors do not bring charges. Why should they? It would be a pointless waste of time; and, moreover, it would constitute harassment of the accused, for which they could be disciplined by the court. Similarly, in clear cases of mercy-killing, where there is no doubt about the patient's hopeless condition or desire to die, charges would not be brought about for the same reasons.[10]

Someone who clings to some of the negative arguments that we've been looking at will quip that inhibitions against euthanasia would come to be so reduced that prosecutors would rarely try to press charges. So, some people might worry about whether Rachels's proposal really avoids the second horn of the dilemma.

3. The Liberty Argument for the Legalization of Voluntary Euthanasia

Let's turn at last to a positive argument for voluntary active and passive euthanasia that Rachels calls the *Argument from Liberty*. The idea behind this argument is that while a person should not be free to do what he wants with the lives of others, he surely ought to be free to do what he wants with his own life (to the extent that this would have no injurious effects on others), and so ought to be free to choose to be euthanized. More formally:

> 10. Everyone has a right to do what s/he wants with his or her own life, provided that this wouldn't harm others in any way.
>
> 11. Legally banning voluntary euthanasia would infringe upon this right.
>
> 12. Therefore, voluntary euthanasia ought not to be legally banned.

This seems like a pretty decent argument.

It also provides a solution to a puzzle I raised earlier stemming from the fact that irrational preferences shouldn't generally be respected. Someone might maintain that irrational preferences that would have bad effects on others deserve no respect and we have no obligation whatsoever to see to it that they aren't frustrated, but concede that irrational preferences that would only have an effect on the life of the person with those preferences ought not to be deliberately frustrated, at least in most cases. There does seem to be a principled asymmetry here, and it is one serious reason why there should indeed be a distinction in permissibility between voluntary, non-voluntary and involuntary euthanasia.

---

[9] Ibid., p.185.
[10] Ibid., p.186.

I still worry a bit about cases in which someone is very, very badly mistaken about the value of his life and seeks euthanasia hastily. Perhaps what we should say about this case is that the person forfeits his right not to have the desires that only affect his life be interfered with.

1. A Little More on the Consequentialism/Deontology Divide

Last Friday I talked at greater length about the distinction between consequentialist and deontological theories. I spent an awful lot of time talking about different versions of consequentialism, but too little time talking about some of the different reasons why people are drawn to deontology. I thought I'd add a few refinements to that discussion so that people don't end up with a distorted or limited perspective on deontological ethics.

As I was saying, most deontologists endorse the following principle:

> *Weak Deontological Dictum* (DD). For at least some acts A and circumstances C, it can be permissible to perform A in C even if some alternative A* to A has substantially better (actual or expected) consequences in C, and even if the general policy of A*-ing in C-like circumstances would, if adopted at large, lead to substantially better (actual or expected) consequences.

A main motivation for deontological ethics that I listed was a desire to accommodate gut intuitions about cases like *Fat Man on the Bridge* in which someone's rights are violated to achieve an optimal outcome. Crucially, that sort of case actually motivates a stronger claim than WDD:

> *Stronger Deontological Dictum* (DD+): For at least some acts A and circumstances C, it can be impermissible to A even if A has substantially better (actual or expected) consequences in C than every alternative, and even if the general policy of A-ing in C-like circumstances would, if adopted at large, lead to substantially better (actual or expected) consequences than every alternative.

This is because most people think that you wouldn't merely be *permitted not* to throw the fat man off the bridge, but that you would *not be permitted* to throw him off the bridge.

Importantly, not all deontologists are willing to slide from DD to DD+, even though DD+ seems necessary to fully explain gut reactions to cases like *Fat Man on the Bridge*. Indeed, some deontologists are motivated by different considerations than cases like *Fat Man on the Bridge* and principles like the Mere Means Principle. Recall that I said that one complaint about maximizing act consequentialism is that it's overly demanding, and that this is why some people turn to satisficing act consequentialism. It's arguable that plenty of versions of satisficing act consequentialism are still overly demanding. One can bring this out by considering the intuitive thought that if a billionaire fails to donate a very substantial portion of his income to charities like Oxfam, he really isn't even doing enough good on impartial grounds: he could save many, many lives and not even have to sacrifice a very decent standard of living by doing this, and so it seems clear that the act of donating is very substantially better than the act of holding onto the money and not really doing anything of significance with it. Some deontologists think that the fact that it seems very odd to say that it's *morally impermissible* for the billionaire to fail to donate a very large portion of his income to charities is a further reason for endorsing DD.

This is obviously a very different reason for endorsing DD. This consideration is often stated more generally as follows:

> *Agent-Centered Prerogatives*. Sometimes performing the impartially best act or even the impartially good enough act would be very demanding to the agent, and may interfere with the unity and integrity of his life. In these cases, there are *agent-centered prerogatives* that give agents the right to choose to refrain from performing the impartially best or impartially good enough act, even though it may still be permissible to perform that act.

Agent-centered prerogatives are arguably built into certain ordinary normative concepts. Most people believe in such a thing as *supererogation* – i.e., in *going beyond the call of duty*. Mother Theresa, for instance, spent most of her life going beyond the call of duty, and did things that, though amazingly good, were not morally required of her. The very idea of a supererogatory act seems to presuppose the coherence of (weak) agent-centered restrictions. After all, a supererogatory act is clearly *significantly morally better* than a merely dutiful alternative; a merely dutiful alternative may clearly not only fail to be the best act, but fail to get even close to being the best act.

The attempt to motivate DD by the existence of agent-centered prerogatives is to be contrasted with the earlier consideration, which is stated more generally as follows:

> *Agent-Centered Restrictions*. Sometimes performing the impartially best act or even the impartially good enough act would require one to do certain intuitively objectionable things to some small class of people – e.g., to use the fat man as a mere means to the impartially optimal end of saving seven lives. In these cases, there are *agent-centered restrictions* that prevent agents from performing the impartially best or impartially good enough act, and that make it impermissible to perform that act.

As I said in less technical language before, agent-centered restrictions seem to be needed to make full sense of *Fat Man on the Bridge*.[11]

Even so, some deontologists are willing to follow act consequentialists in rejecting agent-centered restrictions while departing from them in accepting prerogatives. This position is neither obviously incoherent nor ill-motivated. Why? Well, although they may seem to be needed to capture some gut reactions, restrictions give rise to paradoxes to which prerogatives do not. A full-fledged defender of restrictions will claim that it's wrong to murder one not just to prevent several other *deaths* or *lettings-die* from occurring (as in *Fat Man on the Bridge*), but also to prevent several other *murders* from occurring. This is a very peculiar claim. What is it about the fact that *I* am the murderer of the one that makes the state of affairs in which I murder one and no one else murders *objectively morally worse* than the state of affairs in which I murder none and someone else murders four? Some philosophers – e.g., Samuel Scheffler – think that there is no good answer to this question, but still feel compelled by considerations of personal integrity and demandingness to accept prerogatives.

In any case, the upshot is that we can distinguish three versions of deontological ethics:

---

[11] The terms "agent-centered prerogative" and "agent-centered restriction" were introduced by Scheffler (1982).

> *Weakest Deontology*: There are strong agent-centered prerogatives (e.g., permissions to do acts that aren't even good enough by impartial lights) but no agent-centered restrictions.
>
> *Middling Deontology*. There are agent-centered restrictions but no strong agent-centered prerogatives.
>
> *Strongest Deontology*. There are agent-centered restrictions and strong agent-centered prerogatives.

Kant and his followers generally accept either Middling or Strongest Deontology. But there are some – e.g., Scheffler – who have toyed with Weakest Deontology.

2. Advance Directives

An issue that came up in passing in one of Jeff's lectures that is worth discussing at greater length is the ethical authority of *advance directives*. The main example of an advance directive is a case in which a person makes out a living will that instructs caretakers to do or fail to do certain things under the condition that the person permanently loses his ability to make competent decisions. Someone might write in a living will: "If I contract a terminal illness when I become severely and irreversibly demented, please do not provide me with life support." How a caretaker ought to respond to an advance directive is a relatively straightforward matter when the later stage of the person who made the directive does not oppose it and following through with the directive would really be better for the later stage of the person. It becomes much harder, though, to figure out how a caretaker ought to respond to an advance directive when either (i) the later stage of the person who made the directive wants to live, or (ii) following through with the directive would not be better for the later stage of the person.

Let's focus on a concrete example to make the dialectic a little easier to follow. In an earlier lecture, Jeff brought up the case of Iris Murdoch, a great Oxford philosopher/novelist who ended up with a severe case of Alzheimer's late in her life that obliterated her philosophical and literary abilities, but left her with the ability to enjoy such lowly things as the Teletubbies. Now, as it happens, Iris didn't write an advance directive. But suppose that she had written one, which said something like this: "In the event that I contract a severe case of Alzheimer's that makes me unable to remember my philosophical and literary achievements, and that leaves me with the intellectual capacities of a three year-old, do not keep me on life support if I contract a life-threatening disease." Suppose, in particular, that she wrote this directive because she valued her philosophical and literary abilities above almost everything else in her life, refused to identify with any possible later self-stage that was stripped of these abilities, and regarded the kind of later life that she was in fact fated to briefly live as degrading. Finally, suppose that her later self did contract a fatal disease but could have been kept alive with the aid of life support for an extra year, that this later self-stage would have still gotten enormous pleasure out of simple-minded activities like watching the Teletubbies, and would have requested to be kept on life support, no longer recalling the advance directive that she had originally made.

How should Iris's caretakers act in this circumstance? To answer this hard question, we have to answer the following four easier subordinate questions:

Q1. Does Iris's self-stage fall in the realm of respect? Or is this self-stage morally more comparable to a very young child, late-term fetus or animal that doesn't command the kind of respect that Iris's earlier self commanded?

Q2. Would being kept on life support really be better for Iris's later self-stage?

Q3. Would Iris's life *viewed as a whole* be worse if she were kept on life support rather than being allowed to die?

Q4. Does the reason to take Iris off life support provided by the respect demanded by her younger self's autonomous preference outweigh whatever moral reason there is to keep Iris on life support provided by the fact that being kept alive might be good for this later self-stage?

The answer to the first question in Q1 is arguably 'no', and the answer to the second in Q1 is arguably 'yes'. For, as Jeff has suggested, it seems to be a necessary condition for falling within the realm of respect that one be rational, autonomous, and self-conscious, and Iris's later self-stage possesses none of these properties to any significant degree. Of course, though, this is quite compatible with our having some obligation to see to it that the interests of Iris's later self-stage be promoted, just like how, though animals are not within the realm of respect, the promotion of their interests should figure into our moral reasoning.

The answer to Q2 is probably 'yes', given the way in which the case has been set up. One *might*, however, try to argue for a negative answer to Q2 by giving a negative answer to Q3. In particular, one might reason as follows:

A. What is best for a person at any time in his life is whatever will make that person's life as a whole be as good as possible.

B. Keeping Iris on life support would not be good for her life as a whole, since it keeps her in a state that provides a very humiliating, degrading and sad end to an illustrious and rich intellectual career.

C. So, it is not best for Iris's later self-stage to be kept on life support.

There's an analogy for this argument. A sundae and a piece of pizza might individually be very good things to eat. It doesn't, however, follow that the result of smashing them together would be a very good thing to eat. So, although the sundae may be a very good thing by itself, it may be best to avoid dumping it onto the piece of pizza, and even better to simply have no sundae at all if the only way to have it is to have it as a cool topping for a pizza. According to the A-B-C argument, the same goes for spans of life experience. Although it may be very good and enjoyable for Iris's later self-stage to undergo the experiences that it undergoes, it may not be so great to add it on as a conclusion to Iris's life, and so it may be a better idea to prevent it from occurring at all if it could only occur as the concluding part of her life.

The main problem with this reasoning is that because there is so little psychological continuity between Iris's final self-stage and her earlier self, it makes little sense to endorse (A) or even, for that matter, to think that there is a sensible way in which Iris's life can be

viewed as just one whole that can be an object of clear evaluation. Given the lack of psychological continuity between the two parts of her life, it makes more sense morally to think of there as simply being two different lives to which two different standards of evaluation apply. This is related to the seemingly correct verdict about a case that I brought up earlier in the semester. In that case, you were presented with three options: (i) death, (ii) continuing your life as it is, except while getting 5% less pleasure out of the experiences that you have, or (iii) having your memories, desires, intentions and beliefs completely altered, but also getting vast sums of money that would allow you to lead a really excellent life, as well as a pleasure-enhancement chip installed in your brain. Most people would pick (ii), because the vast psychological changes that you would undergo in (iii) would be relevantly comparable to simply biographically dying. And if you did pick (iii), the person you would become would have no reason to care about the plans and desires of your earlier self, since that earlier self is relevantly like someone else. The same applies here: the later Iris shouldn't care about preserving the kind of unity and quality of life that would have only made sense to her earlier self any more than she should care about the unity and quality of the life of someone else who may get enjoyment and satisfaction out of totally different things.

So, it seems better to say 'yes' rather than 'no' to Q2. Nevertheless, a closely related question seems to deserve an opposite answer. Suppose we ask:

> Q5. Would being kept on life support in a demented state be better for Iris's earlier self?

If we accept the idea that there can be posthumous harms, we ought to regard this as a perfectly intelligible question. Assuming it's intelligible, the answer to Q5 would seem to be 'no'. Being kept on life support harms Iris's earlier self in roughly the same way in which having one's dead body defecated upon and thrown to some sharks would be a harm to one if one previously expressed a wish to be buried clean and intact in a nice graveyard, and paid to have this done.

If this is right, giving an answer to Q4 and to our original question becomes difficult. We have to weigh the "posthumous" harm that would be done to Iris's earlier self against the harm that would be done to Iris's demented self if she were taken off life support. I myself don't have a clear intuition about which harm is greater. Jeff does: he thinks the answer to Q4 is 'yes' and that Iris's later self ought to be taken off life support. Here's what he says:

> I believe that, in this situation, the Demented Patient's physicians ought to give priority to the earlier part of her life, or to her earlier self. They ought, in other words, to allow her to die…. Unlike Dworkin, I do not believe that the Demented Patient's present good is determined by what would be best for her life as a whole. It is, on the contrary, in her present time-relative interest to continue to live. But her continuing to life would, as Dworkin rightly emphasizes, be worse for her life as a whole, which in turn would be worse for the earlier life, making it a component or constituent of a lesser whole. Because the earlier part of the life is overwhelmingly the dominant part, its good should have priority. The earlier part was, in itself, a reasonable full and complete life with its own deep…unity. It was the life of the individual in her higher state, when she was a rational and autonomous person. Its good – which is the good of that earlier, higher self – is therefore more significant than the good of the shallow and necessarily rather brief period of dementia

> dangling at the end of the life…. So the Demented Patient's present good ought to be sacrificed for the greater good of her earlier self, which is also the greater good of her life as a whole.[12]

Thoughts? My intuitions just aren't firm enough to conclusively assess this reasoning.

3. The Cons of Legalized Euthanasia

Even if one accepts that active and passive voluntary euthanasia are morally permissible, one needn't automatically accept that it would be a good idea to legalize them, at least without some considerable restrictions. There are many reasons why one might be inclined to draw a boundary line here, but most of them fall into one of the following categories, each of which I'll discuss at greater length in a moment:

> 1. *The Slippery Slope.* Some worry that if we legalize clearly morally acceptable forms of euthanasia, we will set foot on a slippery slope that will take us to legalizing morally unacceptable forms. Maybe there will be no barrier that will prevent us from eventually legalizing involuntary euthanasia, or from mistakenly regarding as "euthanasia" what may really be worse for the patient.
>
> 2. *Possibility of Abuse.* Some worry that any law that is robust enough to allow all morally permissible cases of euthanasia will be gappy enough to permit at least some morally unacceptable cases. In other words, there might be loopholes in any law that would be strong enough to permit the good cases of euthanasia that could be *abused.*
>
> 3. *Possibility of Mistakes.* Some worry that if we legalize morally permissible types of euthanasia, we might make hasty and mistaken judgments about whether particular cases exemplify these types. There might be cases of misdiagnosis, cases in which someone might have unexpectedly recovered, or cases where a cure might be discovered right after the person is killed. We wouldn't want any of these things to happen, and some people think this is enough to constitute a case against the legalization of euthanasia even under highly restricted conditions (e.g., in which the patient is believed to have a terminal illness and to be incurably suffering to a degree that truly makes life intolerable).
>
> 4. *Bad Side-Effects: Pressure.* Some worry that legalizing clearly morally permissible forms of euthanasia will have adverse side-effects. Perhaps people will end up feeling pressured into giving consent to be euthanized when, deep down, they would prefer not to be euthanized. More specifically, if we ended up accepting some list of conditions under which a life would be regarded as no longer worth living, a patient who constitutes a financial drain or an emotional burden on his family may feel pressured to take himself off the scene even though he doesn't really agree that his life is no longer worth living.

In my substitute lecture for Jeff last Friday, I talked about (3) and (4), and so I'll mostly set these arguments aside unless people want to talk about them. But I think it would be worthwhile to focus a bit more on (1) and (2), since Jeff sped through these arguments pretty quickly.

---

[12] McMahan (2002: 502 – 3).

2.1. *Slippery Slope Arguments*

The slippery slope argument is perhaps the most common one raised against the legalization of euthanasia. There are two unequally plausible forms that this argument sometimes takes:

- *Psychological Slippery Slope.* If we legalize morally permissible practices like voluntary passive and active euthanasia, we will let down our inhibitions against involuntary euthanasia and against killing people whose lives we may mistakenly think aren't worth living, and will end up legalizing these bad practices, too. Since we shouldn't legalize these bad practices, we shouldn't legalize any kind of euthanasia at all.

- *Logical Slippery Slope.* If we legalize morally permissible practices like voluntary passive and active euthanasia, there will be no principled logical reasons not to legalize involuntary euthanasia and the killing of people whose lives we may mistakenly think aren't worth living. So, if we want to have a logically consistent law, we couldn't legalize the good practices without legalizing the bad ones. Since we shouldn't legalize the bad ones, we shouldn't legalize euthanasia at all.

The second version of the argument is clearly mistaken, since there are straightforward and, indeed, obvious reasons why involuntary euthanasia and the killing of people whose lives really *are* worth living are impermissible that do not apply to voluntary active and passive euthanasia. The first version of the argument is less clearly mistaken, since it relies on a psychological hypothesis that isn't obviously disconfirmed.

Some people think that this hypothesis is clearly confirmed, but for bad reasons. One case that people routinely bring up is that the Nazis' program of mass killing had its origins in something that they called 'euthanasia': they thought there was such a thing as a life not worth living, and the people to whom this thought was initially applied were not members of an ethnic group, but simply the severely and chronically sick. Here there was in fact a clear slippery slope down which the Nazis careened. Is this compelling evidence for the hypothesis on which the psychological slippery slope argument relies?

It isn't, for reasons that James Rachels usefully explains in his book:

> Are we to believe that Hitler and his followers were at first an ordinary group of people who permitted mercy-killing from a sense of compassion? And that this led them, in less than a decade, to be transformed into the monsters of concentration camps? Of course this is not what happened…. Among the Nazis, there was never any thought of killing as a compassionate act for the benefit of suffering terminal patients; indeed, this was not even used as a false excuse when they would lie about what they were doing…[n]or was there ever any thought of securing the permission of the victims. The sterilizations as well as the killings were completely involuntary. Where, then, is the analogy with the real euthanasia movement?[13]

To have a real empirical case for the psychological assumption made by the slippery slope argument, we need an example where voluntary passive and active euthanasia were legalized out of a sense of compassion, and where this eventually led to legally accepted clear cases of involuntary euthanasia or killings that weren't really better for the patients. This hasn't been

[13] Rachels (1986: 177 – 178).

shown. It isn't, for instance, illustrated by the case of the Netherlands, since the farthest they have gotten from voluntary active and passive euthanasia are some cases of non-voluntary euthanasia – cases in which patients were unable to express their preferences, and had expressed no prior preferences about how to be treated in these conditions. And there are perfectly good reasons for thinking that these types of cases can be morally permissible. So, the psychological argument rests on an assumption for which there is no real evidence.

2.2. *Arguments from the Possibility of Abuse*

Arguments against the legalization of voluntary active and passive euthanasia that proceed from concerns about abuse take many forms. One is what we might call the *Argument from Bad but Unknowable Motivations*:

> 1. If voluntary active euthanasia were legalized, there would inevitably be legally permitted cases in which the person who provides the (putative) euthanasia has very corrupt reasons for it – say, he simply wants to get his inheritance and doesn't give a damn for the well-being of the person asking for death – though he very effectively hides those reasons and feigns solicitude.
>
> 2. There shouldn't be legally permitted cases in which the person who provides the (putative) euthanasia has motivations that are this corrupt.
>
> 3. Therefore, voluntary active euthanasia shouldn't be legalized.

Premise (1) may be true. It is hard to see how the law could detect cases in which the (putative) euthanizer secretly only grants the wish of the (putatively) euthanized person for reprehensible reasons, but in which he is careful to prevent any evidence of his corrupt motives from showing up. It is less obvious that the inference from (1) and (2) to (3), or, for that matter, premise (2) works. I suspect that this is a general problem that shouldn't be solved by refusing legal permission: *any* legal setup will contain loopholes that permit someone to perform an act that is permissible with intentions that are reprehensible. Given the distinction between evaluations of acts (permissibility/impermissibility) and evaluations of agents (praiseworthiness/blameworthiness), and the fact that what's really at stake here is the potential legal permission of morally impermissible acts, it's not clear that there is really a serious worry behind the argument. People's motives might be forever inscrutable to us; what really matters is that they never perform an act that is itself impermissible. And the Argument from Bad but Unknowable Motivations does nothing to establish that such acts will be allowed.[14]

---

[14] Admittedly, making use of the act/agent distinction is trickier when we turn from morality to criminal law, since it's arguable that punishment should turn more on whether someone was blameworthy than on whether someone brought about some bad upshot. In some cases of attempted murder, for instance, nothing bad actually happens. The person whom the murderer was trying to murder may not in any way be harmed, and may not even realize that there was anybody trying to harm him. Perhaps, for instance, the murderer was a sharpshooter who was caught by a watchman in a tower just as he was about to fire his gun at some people on the street below. Here nothing bad actually happens, and so, at least on *some* normative theories, no impermissible act has been performed. This would follow, for instance, on a consequentialist theory on which the wrongness of an act turns on whether it had any bad actual consequences. But not everyone will agree that the sharpshooter ought to get away scot free in the imagined case! If so, then we can't assume that the criminal law ought to address only impermissible acts; it also ought to address some *merely blameworthy* acts. So, perhaps the defender of the Argument from Unknowable Bad Motivations would complain about what I've said. Still, I

Another worry about abuse is presented by the *Argument from Badly Persuaded Consent*:

> 4. If voluntary active euthanasia were legalized, there would inevitably be legally permitted cases in which someone is euthanized who came to desire and request death only as a result of illegitimate persuasion, brainwashing or the like.
>
> 5. There shouldn't be legally permitted cases in which someone is euthanized who came to desire and request death only for these reasons.
>
> 6. Therefore, voluntary active euthanasia shouldn't be legalized.

This argument is a little more compelling than the last, since here a case could be made that the law may end up permitting some genuinely impermissible acts. Euthanizing someone who only wants to be euthanized because he's been brainwashed would not seem to show respect for his autonomy, particularly if the people doing the euthanizing are the very same people who did the brainwashing. It is far from obvious that this would not be morally wrong. Here, too, though, I suspect that there may be another completely general problem that could only be fully avoided by an impossibly prohibitive law. A clever ex-rapist might discover ways of brainwashing potential victims into having consensual sex with him, and may be able to prevent any evidence from accumulating that would be sufficient to establish that he was clearly violating his victims' autonomy. One might think that this is wrong, but there seems to be little we can do to design a law that would *absolutely guarantee* that this would never happen. A law that absolutely guarantees prevention of all wrongdoers from going scot free for their wrongdoings may not be feasible. The best one can do is weigh the costs of having a fairly prohibitive law against the benefits of having a more liberal one. And nothing has been shown that the benefits of allowing voluntary active euthanasia, at least in a sufficiently careful fashion, would be outweighed by the costs.

---

think the point about the implausible overgeneralizability of the argument holds: this is a broader problem, and to avoid it would almost certainly require a law that was prohibitive to an extreme and unreasonable degree.

1. Self-Defensive Killing: Some Theories and Distinctions

In thinking about self-defensive killing, it's important to distinguish two questions:

> Q1. Under what conditions would some innocent agent A be morally permitted to kill some other agent B who poses a threat to A?
>
> Q2. Under what conditions would some innocent agent A be morally blameless for killing some other agent B who poses a threat to A?

These questions are clearly conceptually different. And given the substantive distinction we've been drawing between blameworthiness/blamelessness and impermissibility/permissibility, there is virtually universal agreement that they ought to receive different answers. Virtually everyone, for instance, will agree to the following claim:

> *Excusable Killing*: If B poses a very serious threat to some innocent person A's life, A couldn't be blamed for killing B in self-defense if killing B was necessary to avert B's threat.

But not virtually everyone will agree to the following counterpart of this claim:

> *Permissible Killing*: If B poses a very serious threat to some innocent person A's life, A is permitted to kill B in self-defense if killing B was necessary to avert the threat.

Why not? Well, to see why not, it's best to get some theories about Q1 on the table. Here are four increasingly strong theories that answer Q1:

> *The Agent-Centered Account*: Innocent agent A is morally permitted to kill some agent B who poses a threat to A if and only if (=iff) (i) killing B is necessary for averting the threat to A, and (ii) killing B is a proportionate means for averting the threat.
>
> *The Rights-Based Account*: Innocent agent A is morally permitted to kill some agent B who poses a threat to A iff (i) B violates A's rights in posing a threat to A, (ii) killing B is necessary for averting the threat to A, and (iii) killing B is a proportionate means for preventing the rights-violation presented by B's threat.
>
> *The Responsibility Account*: Innocent agent A is morally permitted to kill some agent B who poses a threat to A iff (i) B is responsible for the threat he poses to A, (ii) killing B is necessary for averting B's threat, and (iii) killing B is a proportionate means for averting B's threat.
>
> *The Culpability Account*: Innocent agent A is morally permitted to kill some agent B who poses a threat to A iff (i) B is blameworthy for the threat he poses to A, (ii) killing B is necessary for averting B's threat, and (iii) killing B is a proportionate means for averting B's threat.

To see the ways in which these theories disagree, let's bear the following four cases in mind:

Innocent Threat. A huge man is enjoying a picnic on a cliff directly above the deck on which you are lying with your leg in traction. Suddenly a villain pushes him off the cliff. If he lands on you he will kill you, but he would survive because you would cushion his fall. You can save yourself only by hoisting your sun umbrella and impaling him on it.

Nonresponsible Threat. Someone put a hallucinogen in Bob's coffee. Bob looks at you, and hallucinates that you are a big bear that is about to rip him to shreds if he doesn't shoot it in the head. You have no time to do anything to avert his threat except to whip out your flame thrower and scorch him to death.

Justified Threat. You have been tightly chained down to a railroad track by a maniac, but discover a gun next to your hand (which is free) after the maniac has left. On another track, seven other people have been chained down by the same maniac. A trolley is rapidly approaching a junction at which it could turn either onto your track or the track containing the seven others; if it isn't interfered with, it will go down the latter track. A bystander notices what is about to happen, sees a switch that controls the junction, and plans to pull the switch so that it will kill just you instead of the seven. Unless you shoot him dead, he'll still pull the switch.

Culpable Threat. A cold-blooded killer is about to lunge at you and slit your throat. You are no good at hands-on combat, and the only thing you've got to avert the threat is a flamethrower.

The four theories we just stated make the following verdicts about the moral permissibility of your engaging in self-defensive killing these cases:

Threats/Theories	*Agent-Centered*	*Rights-Based*	*Responsibility*	*Culpability*
Innocent	Permissible	Impermissible (?)	Impermissible	Impermissible
Nonresponsible	Permissible	Permissible (?)	Impermissible	Impermissible
Justified	Permissible	Permissible (?)	Permissible	Impermissible
Culpable	Permissible	Permissible	Permissible	Permissible

I'll turn in a moment to explain the several question marks qualifying the verdicts of the Rights-Based Account. Getting back to our original question, though, you might immediately say: "Surely the Culpability and Responsibility Theories are false. Of course you're permitted to defend yourself in *Innocent Threat* and *Nonresponsible Threat*."

But you would speak a little too quickly, since it must be realized that the defenders of these theories could reduce the *prima facie* implausibility of their predictions by digging in their heels about the distinction between impermissibility and blameworthiness. Defenders of these theories could say: "Don't get me wrong: I don't think you would be even slightly blameworthy for defending yourself in these cases. All I'm saying is that, blameless though you may be, the agents in these cases *did not make themselves liable* to self-defensive killing. And to kill someone who hasn't made himself liable to self-defensive killing is always impermissible, though of course sometimes completely excusable." And, on first blush, the defender of either of these theories would seem to have a point. Once this remark and the distinction on which it rests are appreciated, we cannot say that these theories make

obviously false and crazy predictions. It's for this reason that one can endorse *Excusable Killing* and not *Permissible Killing*, and that one can give very different answers to Q1 and Q2.

Having made this point, let me get back to the reasons for the several question marks qualifying the verdicts of the Rights-Based Account. To see the reasons, notice that one could take different stances on the following substantive claims about rights-violations:

1. If Y's rights are violated by X, X intentionally violates Y's rights.

2. If Y's rights are violated by X, X responsibly violates Y's rights.

3. If X infringes Y's rights, X violates Y's rights.

While (1) may seem like a trivial truth, some people are committed to rejecting it – e.g., Thomson. Such people are committed to claiming that a rights-violation occurs in *Innocent Threat*: hence the question mark for that case. (2) is not a trivial truth, but it is supported by the fact that it's far from clear whether the correct description of *Nonresponsible Threat* is that Bob is about to violate your rights *precisely because* Bob is not acting responsibly, given that he's been drugged. One might, however, insist that rights-violation requires only intentional action, not responsible or autonomous action. If one did, one would claim that a rights-violation occurs in *Nonresponsible Threat* and accept the initial verdict listed above for the Rights-Based Account.

(3) is unobvious because lots of people think there's a distinction between infringing a right and violating a right. What a right not to be killed does is provide a *prima facie* reason not to be killed. This reason might be outweighed. If it is, then your right is *overridden*, and so any act that conflicts with your right will only *infringe* that right. Will it also *violate* that right? Well, not obviously, because it's not clear that we can coherently speak of an *overridden right* being *violated*. If you do find it incoherent to talk of an overridden right being violated, you could reject (3). And if you reject (3), you might, if you're a Rights-Based Theorist, accept the verdict already listed for *Justified Threat*. But if you think that we can coherently talk of an overridden right still being violated, you'll accept (3) and, if you accept the letter of the Rights-Based Account stated above, you'll have to revise the initial verdict about *Justified Threat* and claim that it's actually impermissible to engage in self-defense in this case (though perhaps blameless).

2. Thomson on Self-Defensive Killing

Let's turn to a defense of a Rights-Based Account presented in Judy Thomson's "Self-Defense". Although it's classic, this particular species of the account is a bit nonstandard, since Thomson has a rather nonstandard conception of what is involved in violating people's rights. Thomson is aware that there are alternatives to her conception: she simply seems to lack the intuitions that motivate these alternatives, and doesn't do much to defend her competing intuitions. All the same, there are plenty of valuable points worth absorbing in the paper, not just for seeing what motivates the Rights-Based Account, but for seeing some deeper conceptual truths about self-defense and a number of other related issues.

Thomson starts the paper by considering and criticizing several variants of a much simpler account of the morality of self-defense, which I'll call *the Naïve Account*. The first variant is:

> *Naïve Account I*: It is morally permissible for X to do whatever is necessary (and *only* whatever is necessary) to fend off a threat to his life by a responsible and culpable aggressor.

Naïve Account I is stronger than the Culpability Account we saw above in two respects. Firstly, it says that X can do *whatever* is necessary to fend off the threat, not just that it is permissible to *kill the aggressor* if *that* is necessary to fend off the treat. Secondly, it doesn't incorporate any proportionality constraint like condition (iii) in the Culpability Account.

While Thomson attacks these missing components of Naïve Account I, the first thing she does is to simply say that she doesn't find intuitively acceptable the prediction it entails about cases like *Nonresponsible Threat*. She says the following, doing little to explain her intuition: "[T]o say that self-defense is merely excusable in [*Nonresponsible Threat*] is to say that although you would not be at fault…you act wrongly…. I think that cannot be right. (I think it an excessively high-minded conception of the requirements of morality)."[15] Accordingly, she concludes that Naïve Account I is "stronger than it need have been".[16] She says basically the same sort of thing about the predictions of Naïve Account I in the case of *Innocent Threat*, but this time she's a bit more careful in considering the arguments of the other side.

She starts by saying rather flat-footedly: "I suspect that even more people would say that self-defense is merely excusable in *Innocent Threat*. Why so? The innocent aggressor, though without fault, is at least aggressing against you; the fat man is not only without fault, he is not doing anything at all…. *I* think that difference makes no moral difference, and thus that it is permissible for you to proceed in *Innocent Threat* just as in [*Culpable Threat* and *Nonresponsible Threat*]."[17] She does admit that defenders of a suitable revision of Naïve Account I might insist that the crucial difference just is that there's *killing* in *Nonresponsible Threat* but no killing in *Innocent Threat*. But then she denies that there isn't killing in *Innocent Threat*, saying: "I think we should not be moved by this idea. Is it really to be thought that Y kills X only if Y aggresses against X? Suppose a piano and a safe fell off a roof, and we know that one fell on Alfred, and that the event that consisted in its fall on him killed him. We might ask, 'which killed Alfred, the piano or the safe?' The correct answer might be. 'The piano', despite the fact that pianos commit no acts of aggression…. Indeed, I should think that if an event that consists in the fall of Y on X kills X, then it follows that Y killed X, whatever Y may be."[18] So, she concludes that Naive Account I was too strong in another respect.

But she does not think we should conclude from this that the following is a good theory:

> *Naïve Account II*: It is morally permissible for X to do whatever is necessary (and *only* whatever is necessary) to fend off a threat to his life by any Y who threatens him, even if Y is nonresponsible and blameless.

---

[15] Thomson (1991: 285).
[16] Ibid., p.286.
[17] Ibid., p.287.
[18] Ibid., p.289.

Her reason for rejecting Naïve Account II is quite simple: it is way too weak on several scores. Here are three cases she presents in which it makes seemingly incorrect predictions:

> *Substitution of a Bystander.* A villain has started a trolley down a track toward you. You cannot stop the trolley, but you can deflect it. Unfortunately, the only path onto which you can deflect it will take it onto a bystander who cannot get off the path in time.
>
> *Use of a Bystander.* A villain has started a trolley down a track toward you, and the only way you have of defending yourself is to shoot a bystander who stands on a footpath over the track:. He is sufficiently heavy to crush the trolley's rooftop mechanism when he falls onto it, which will thereby stop the trolley.
>
> *Running Roughshod over a Bystander.* A villain is shooting at you, and your only defense is to run. But your only path to safety lies across a bridge that will hold only one person, and there is already a man on it; if you rush onto the bridge, he will be toppled off it into the valley below. [19]

Naïve Account II entails that it would be permissible for you to do the controversial acts in question in all three of these cases. Thomson thinks that this is obviously wrong.

She concedes that *Use of a Bystander* might look a bit worse to some people, since the controversial act in this case would seem to violate the Mere Means Principle we discussed last week in a way that the other acts wouldn't. But she thinks there's no clear moral difference in permissibility between killing someone as a *necessary side-effect* of one's act and using him as a *mere* means. As she explains in the following interesting and not entirely obvious passage:

> Why, after all, should it have been thought that the fact that you *need* the person you would have to kill in order to defend yourself makes it worse for you to proceed than it would have been if you had not needed the person? If I am right in thinking that this is the mark of a Use of a Bystander case, then *using* a person does not in general have the special moral taint that has been ascribed to it. Appeals to the notion of 'respect for persons' will certainly not suffice to make out this special moral taint. After all, if one proceeds in a Substitution of a Bystander case…or in a Riding Roughshod over a Bystander case, one behaves as if the person one kills *were not there at all* – surely no less a display of a lack of respect for persons.[20]

This seems a little quick.

But she ultimately appreciates this fact in her discussion of the *Doctrine of Double Effect*, the thesis according to which (in Thomson's formulation) "we may do what will cause a bad outcome in order to cause a good outcome if and only if (1) the good is in appropriate proportion to the bad *and* (2) we do not intend the bad outcome as a means to the good outcome."[21] This doctrine strikes a lot of people as intuitive. It seems to explain, for

---

[19] I take these *verbatim* from Ibid., pp.289 – 290.
[20] Ibid., p.291.
[21] Ibid., p.292.

instance, why it might seem permissible to kill some civilians as a side-effect of an otherwise fully justified bombing.

Thomson, however, thinks that the doctrine is at least partly confused, and probably isn't ultimately needed to explain our intuitions about this kind of case. Her reason for thinking that it's confused is that she doesn't think intention should have any effect on permissibility, though she grants that it ought to have some effect on blameworthiness. I think she's right about this for reasons that should be obvious from previous classes – i.e., I think there's a large gulf between impermissibility and blameworthiness, and that properties of the agent's mental life almost always only affect the latter without affecting the former. She does an even better job than I've probably done in arguing for this claim in the case of intention:

> Suppose a pilot comes to us with a request for advice: "See, we're at war with a villainous country called Bad, and my superiors have ordered me to drop some bombs at Placetown in Bad. Now there's a munitions factory at Placetown, but there's a children's hospital there too. Is it permissible for me to drop the bombs?" And suppose that we made the following reply: "Well, it all depends on what your intentions would be in dropping the bombs. If you would be intending to destroy the munitions factory and thereby win the war, merely foreseeing, though not intending, the deaths of the children, then yes, you may drop the bombs. On the other hand, if you would be intending to destroy the children and thereby terrorize the Bads and thereby win the war, merely foreseeing, though not intending, the destruction of the munitions factory, then no, you may not drop the bombs." What a queer performance this would be![22]

This seems pretty decisive to me. Of course, it leaves open the question of why it *is* intuitively permissible to kill civilians as a side-effect of a fully justified bombing in a just war. But I, like Thomson, will set this question aside until we get to discussing the morality of killing in war.

In any case, the upshot of all this was that Naïve Account II fails. Thomson's suggested replacement for it is a version of the Rights-Based Account. She thinks that what accounts for the asymmetry in permissibility between cases like *Use of a Bystander* and *Nonresponsible Threat* is that (a) the fact that your life is threatened by the killer gives you a right to kill him in self-defense *because* he has lost his right not to be killed by so threatening you (and since he never had a right to kill you in the first place), and (b) the innocent bystanders have not lost their rights not to be killed by you, and your right to defend yourself does not turn into a right to use them as a means of fending off the threat, even if it's necessary.

Of course, the controversial bit in all this is the idea that the people threatening you in *Innocent Threat* and *Nonresponsible Threat* really are violating your rights, and the connected idea that they have lost their rights not to be killed by you. Thomson admits that there could be some dispute about this, but she's content to simply state her competing intuitions and conclude that, whatever we might say about these ideas, the Rights-Based Account will stand. As she puts it:

---

[22] Ibid., p.293.

> [W]hat makes it permissible for you to kill the two drivers and the fat man is the fact that they will otherwise violate your rights that they may not kill you, and therefore lack rights that you not kill them. Some people may say there is no such fact in the case of either the fault-free [aggressor] or the fat man…. All is well for the account of self-defense I have offered if the first group are also content to say that it is impermissible (though excusable) for you to proceed in [these cases].[23]

In other words, this is just an in-house dispute, and what matters most to her is the success of the skeletal structure of the theory, not the particular view of rights-violation she favors.

Thomson wraps up her paper with a couple of further claims. One is what I'll call:

> *The Asymmetry Thesis*: If it is permissible for X to kill Y in self-defense, then it is impermissible for Y to fight back.

The Asymmetry Thesis seems plausible enough in cases like *Culpable Threat*. It's obviously crazy to think that if a murderer is about to kill you, and you start preparing to defend yourself, it's suddenly morally permissible for him to kill you because he's now got to defend himself. Even Hobbes would agree with this claim, since, while he would think that the murderer has a *right* to fight back, it is not a *moral* right. But the Asymmetry Thesis does not seem plausible for cases like *Innocent Threat*. Thomson disagrees, but this strikes me as implausible and unsupported. Surely the Fat Man would be morally permitted to force your sun umbrella out of the way if there were some way he could do this without impaling himself on it. But perhaps what this shows is not that the Asymmetry Thesis is false, but simply that Thomson is wrong that you're morally permitted to kill in self-defense here: you're merely fully excusable.

Another intriguing thesis that Thomson suggests in wrapping up her paper is what I'll call:

> *The Agent Neutrality Thesis*: If it is permissible for X to kill Y in self-defense, then any third party Z would be permitted to kill Y to defend X.

As with the Asymmetry Thesis, it's clear that the Agent Neutrality Thesis is correct in cases like *Culpable Threat*. But, if we agree with Thomson that you're morally permitted to defend yourself in *Innocent Threat*, we may have to reject the Agent Neutrality Thesis. For it isn't so clear that, if a bystander with a bazooka saw the fat man falling towards you in a variation of the case in which you don't have the umbrella, and blew the fat man to bits before he fell on you, thereby saving your life, the bystander would be doing something morally permissible. But maybe some people would be willing to embrace this conclusion. If not, then it looks like we either have to abandon the Agent Neutrality Thesis or give up on the claim that you're morally permitted to kill in self-defense here: you're merely fully excusable.

3. Otsuka on Self-Defensive Killing

Let's now turn a view that departs from Thomson's in small but important ways. In his "Killing the Innocent in Self-Defense", Michael Otsuka defends a version of the Responsibility Account that also qualifies as a restricted version of the Rights-Based

[23] Ibid., p.303.

Account. Otsuka's view is exactly like Thomson's except that he denies that the threatening persons in *Innocent Threat* and *Nonresponsible Threat* lose their right not to be killed in self-defense. He thus holds that it is impermissible, though fully excusable, to kill innocent and nonresponsible threats.

He starts the paper with a defense of the first half of the claim that has the following form:

> P1. It is impermissible to kill a bystander to prevent oneself from being killed.
>
> P2. The killing of an innocent threat and the killing of a bystander are, other things being equal, on a par as far as permissibility is concerned.
>
> C. So, it is impermissible to kill an innocent threat (though maybe fully excusable).[24]

He spends a bit of time explaining why we ought to accept P1, but his reasons are pretty much the same as Thomson's, so I think we can skip this part of the paper. He then proceeds to offer indirect and direct arguments for P2. The indirect arguments really amount to rebuttals of the most salient objections to P2.

One objection he considers rests on an appeal to the supposed moral difference between killing and letting die. Since the innocent threat will otherwise kill you while the bystander will not, one might think that the innocent threat is more liable to killing than the bystander. Otsuka's reply is this: "I believe that the moral distinction between killing and letting die gains plausibility from particular facts about the wrongness or moral badness of killings that are, or are the consequence of, intentional acts of morally responsible agents, especially when those acts are intended or foreseen as killings. It is natural to think that one bears special moral responsibility for such killings that flow from one's agency. No such special responsibility, however, plausibly attaches to lethal movements of one's body over which one has no control."[25]

The second objection he considers follows from Thomson's view of the cases: the bystander will not violate your rights, but the innocent threat will. This difference, according to Thomson, grounds the difference in permissibility. Otsuka's reply to this objection is the most obvious one: "I…reject the premise that, if they kill you, [Innocent] Threats will violate your right not to be killed."[26] He strengthens this reply by noting that it is utterly implausible to think that if a tornado whisks a fat man into the sky and drops him so that he's about to fall on you unless you whip out your sun umbrella and impale him, he is about to violate your rights. Since there's no relevant difference between this case and the original *Innocent Threat*, the objection fails.

Otsuka notes that one could insist that while the fat man falling from the tornado's clutches does not *himself* violate your rights, he does cause your rights *to be violated*. But he replies that he is "no more convinced by the claim that a naturally falling stone such as a meteorite can cause a rights violation than…by the claim that such an object can violate a right".[27]

---

[24] Otsuka (1994: 76).
[25] Ibid., p.79.
[26] Ibid., p.79.
[27] Ibid., pp.81 – 82.

A final objection he considers follows from the idea, favored by some defenders of the Agent-Centered Account, that people enjoy a certain kind of sovereignty over their own 'space', which gives them a right to eject from their space whatever finds itself there without their autonomous consent. Otsuka's reply to this idea is that while it may be correct, the rights-violation provided *just* by the invasion of space is insufficient to justify self-defensive killing, and so this point does nothing to advance the case against P2:

> Imagine, for example, that a small child has somehow been handcuffed and strapped to your body and will remain attached for the next several minutes. Whether or not such an invasion of your space constitutes a rights violation, surely you may not vaporize the child even if that is the only way to stop her from occupying your space. At best, you could respond in such fashion only if your being attached to the child were killing you or causing serious injury. *But now the fact that the child will kill…you grounds the claim that you may kill her, and not the fact that she will invade your space. [But] I have…already argued that such a person could not be said to violate your right not to be killed if she kills you in this fashion. The mere fact that a[n] [Innocent] Threat, but not a Bystander, will kill you unless you kill her is not enough to differentiate a threat from a Bystander and justify the killing of the former but not the latter.*[28]

Having rebutted these three objections, Otsuka presents his positive argument for P2. To grasp the argument, we need to introduce three cases on which it relies:

> *Bystander Near the Tracks.* An innocent person is lying along side the path of a runaway trolley car. Unless you hurl at that trolley a bomb that you know will also kill the innocent person, the trolley will run you over.
>
> *Innocent Person Trapped in the Trolley.* An innocent person is trapped inside a runaway trolley car. Unless you hurl a bomb that will destroy the trolley, and hence also the innocent person, the trolley will run you over before coming to a gentle stop.
>
> *Falling Person with Huge Ski Boots.* An innocent person has been thrown off a cliff. The villain who threw him off the cliff attached gigantic ski boots to his feet – boots so big that they (and they alone) will kill you, who are down below, unless you fire a rocket at the guy.[29]

Otsuka's positive argument for P2 seems to be this:

1. It is impermissible to hurl the bomb at the trolley in *Bystander Near the Tracks*.

2. If it is impermissible to hurl the bomb at the trolley in *Bystander Near the Tracks*, then it is impermissible to hurl the bomb at the trolley in *Innocent Person Trapped in the Trolley*, since the only difference between the cases is the spatial location of the person, and that's not a relevant difference.

3. If it is impermissible to hurl the bomb at the trolley in *Innocent Person Trapped in the Trolley*, it is impermissible to use your sun umbrella to impale the fat man in *Innocent Threat*.

---

[28] Ibid., p.83.

[29] These cases are taken pretty much verbatim from Ibid., pp.85 – 86.

a. Argument for (3): The only reason why you would deny (3) is that you think that there's some moral difference between the fact that the fat man's *body* kills you in *Innocent Threat* and the fact the *trolley* kills you in *Innocent Person Trapped in the Trolley*. But if you really buy this reasoning, and you think it's impermissible *only* for this reason to blow up the trolley in the case at issue, you are also forced to claim that it is impermissible to fire the rocket at the innocent man in *Falling Person with Huge Ski Boots*. But there is no intuitive difference between *Falling Person with Huge Ski Boots* and *Innocent Threat*. So, either you claim that they're both impermissible or you claim that they're both permissible. Either way, you will lose your argument against (3). Therefore, (3) stands.

4. Therefore, it is impermissible to use your sun umbrella to impale the fat man in *Innocent Threat*.

5. Therefore, *Bystander Near the Tracks* and *Innocent Threat* are morally equivalent.

6. Since these were arbitrary cases, any similar Bystander case will be morally equivalent to any similar Innocent Threat case.

This establishes P2. Otsuka sums up the upshot of this nice argument in these terms:

> Those, such as Thomson, who believe that it is permissible for you to kill the falling Threat but impermissible for you to do that which you know will kill the Bystander lying alongside the trolley, are faced with the following dilemma. If they believe that it is permissible for you to destroy the trolley inside of which an innocent person is trapped, then they must perform the difficult task of explaining why an innocent inside a trolley may be killed, whereas one who is near the trolley may not. However, if they believe that it is impermissible for you to destroy the trolley with the trapped innocent, then they must explain why such an innocent may not be killed, whereas a Threat that is falling toward you may be. My own view is that killing a person alongside the trolley, killing a person inside the trolley, and killing a falling Threat are on a par as far as permissibility is concerned.[30]

This strikes me as a powerful argument.

Otsuka's argument that it is impermissible to kill a *nonresponsible* threat seems a bit sketchier. He asserts without argument the following claim: "I believe…that the presence or absence of harmful agency is morally relevant only in cases involving those who are functioning as morally responsible agents. Yet [Nonresponsible Threats] and [Innocent Threats] are, by hypothesis, not functioning as morally responsible agents. I believe, therefore, that the presence or absence of harmful agency is morally irrelevant in these cases."[31] Given this claim, the argument then seems to be that since killing is impermissible in *Innocent Threat* (which follows from the P1 – C argument), and there is no morally relevant difference between *Innocent Threat* and *Nonresponsible Threat*, killing is also impermissible in *Nonresponsible Threat*. This argument seems a little weak to me, since you might just run it in reverse: there

[30] Ibid., p.86.
[31] Ibid., p.90.

is a difference in permissibility between these cases, and therefore the presence or absence of harmful agency is morally relevant even in the absence of responsibility for that agency.

Otsuka does, however, seem to have a different argument for his conclusion. The argument is that since a nonresponsible threat cannot violate your rights, and the Rights-Based Account is true, you are not permitted to kill a nonresponsible threat. What's the motivation for the first assumption in this argument? Otsuka says this: "An angry grizzly bear on the attack against a human being possesses some of the marks of an Aggressor: it lacks moral responsibility for what it does, and yet it [intentionally] acts in order to harm. Yet there is little temptation to say that a grizzly bear can violate your right not to be killed." Since the grizzly bear does, Otsuka thinks, have the intention to kill you (albeit in a somewhat primitive form), intention to kill alone cannot suffice for rights-violation, and so responsibility is also necessary. This seems fairly plausible.

But this argument falls a bit short of its intended conclusion. One *can* be culpable for one's nonresponsibility on some occasion. Suppose that Bill is normally too nervous to kill people, and so he takes a drug that makes him go temporarily crazy to make it possible for him to kill Dave, whom he despises. When he's drugged, Bill does not qualify as a responsible agent. But since he was responsible, as it were, for his temporary nonresponsibility, he can be held to blame for what he does in this state. More strongly, one might think that he *can* violate someone's rights even in this state in virtue of the fact that his temporary nonresponsibility had a malicious intention as its cause. If so, then at least some temporarily nonresponsible threats could be viewed as liable. If so, it ought to be permissible to kill them in self-defense, and not just excusable. If so, Otsuka's conclusion needs to be weakened just a bit.

1. The Ethics of War: Distinctions and Theories

The majority of theorists who believe in the possibility of just war acknowledge two sets of principles that are relevant to determining how morally acceptable some war is overall:

A. The principles of *jus ad bellum*, which govern the *resort to war*.
B. The principles of *jus in bello*, which govern the *conduct of combatants in a war*.

There are two strikingly different views about the relations between these two sets of principles.

On the widely accepted *Traditional Theory of the Just War*, the principles of *jus ad bellum* and *jus in bello* are completely independent. The upshot of this claim – which I'll call the *Independence Thesis* – is that an unjust war can be justly fought, and a just war can be unjustly fought.

Something that is closely tied up with the traditional view's endorsement of the Independence Thesis is the doctrine of the *moral equality of combatants*, according to which there is just one set of rules for *jus in bello* that apply *equally to all combatants*, whether they are on the justified or the unjustified side. A surprising consequence of this doctrine is that combatants on the unjust side *do no wrong* merely by participating in the war efforts of that side: they only do wrong if they violate the principles of *jus in bello*, which are completely independent of whether the resort to war was itself just. So, the Traditional Theory does not *merely* make the uninteresting claim that, even when a war is unjust, there ought to be a set of *conventions* governing the way in which it is to be fought, and hence that acts in such a war can be *conventionally justified*. It makes the interesting claim that the efforts of combatants on the unjust side can be *as morally justified as the efforts of combatants on the just side*, and that the combatants on the unjust side have all the same undefeated rights as the combatants on the just side.

And that is a very strong claim. Surprisingly, this claim and the Independence Thesis on which it rests have only come under serious dispute in recent years. According to the *Revisionist Theory of the Just War* (centrally defended by Jeff McMahan in his (2004) and (2009)), the Independence Thesis is false: the principles of *jus in bello* depend on the principles of *jus ad bellum*, the efforts of the combatants on the unjust side cannot be as morally justified as the efforts of combatants on the just side, and the undefeated rights of the combatants on both sides are not equal.

I'll turn to part of Jeff's case in favor of the Revisionist Theory in §4. For the moment, it is just worth stressing how intuitive his view ought to seem, and how unintuitive the Traditional Theory ought to seem. Recall that part of the motivation for the very idea of a just war derives from a strong analogy ("the domestic analogy") with individual self-defense: at minimum, if one state is responsible for an unjust attack on another state, the latter state ought to be just as justified in fighting back as an individual person would be in response to an unjustified and culpable attack from another individual person. The morality of self-defense is notably asymmetrical. As we saw, it is ludicrous to think that if a cold-blooded

killer pulls a knife on you and runs at you, and you pull out your knife and jab him, he is suddenly now justified in attacking you. If we accept the domestic analogy, we ought to think the same thing at the level of states: when the just state responds to an unjust attack with acts of violence, we shouldn't think that it is *now* morally fine for the unjust state to strike back. Curiously, the Traditional Theory accepts this claim, but *then* goes on to deny that it implies that the individual combatants on the unjust side would be morally unjustified in contributing to these attacks. This is a *prima facie* strange thought, and defenders of the Revisionist Theory think it is flat-out incoherent.

2. Principles of *Jus ad Bellum*

Let's set aside the disagreement between the Traditional and Revisionist theories for the moment and look at some neutral territory: the principles of *jus ad bellum*. Many just war theorists on both sides agree that all or most of the following are necessary conditions for the resort to war to be morally permissible, each of which I'll discuss at some length:

I. Just Cause
II. Competent Authority
III. Right Intention
IV. Necessity ("Last Resort")
IV. Proportionality
VI. Reasonable Hope of Success

As we'll see, there are reasons to doubt that conditions II, III and VI are really necessary.

2.1. *Requirement of Just Cause*

While all agree that I is a necessary condition for a just war, there is much disagreement about what causes are just. International law is highly restrictive in what it recognizes as a just cause: the United Nations Charter holds that it's illegal for one state to attack another except when the latter has in fact used an armed attack against the former. A parallel understanding of a morally (and not just legally) just cause would hold that it is permissible for one state to go to war against another if and only if the latter state has committed a serious wrong against the former. Depending on exactly what we count as a "wrong", this condition might go beyond the condition recognized in the United Nations Charter. If the relevant type of wrong is an act of aggression, the two coincide; if the relevant class of wrongs also includes such things as unjustified seizings of land, the two needn't coincide.

But however the class of wrongs is circumscribed, this view is *overly restrictive* in several ways. Firstly, entails that it's impermissible for a state to go to war against another if the former is guilty of serious, culpable wrongs to its own people. Secondly, it rules out the permissibility of one state's defending *another* state from unjust and culpable attacks by some third state. Thirdly, it rules out the permissibility of preventive attacks against *very likely* threats that have not actually occurred. Many would want cases in these categories to count as permissible.

Less obviously, the narrow view we were considering at first is also *overly permissive*, at least if it isn't qualified in certain ways. To see this, we have to engage in a bit of science fiction. Suppose that a mad scientist installs mind control chips in the heads of all the officials of some state A, and directs them against what would otherwise have been their wills to prepare

a nuclear strike on some unsuspecting state B. If we cling to the distinction between assessments of acts and agents, we can sensibly say that A wrongfully attacks B. Nevertheless, A doesn't *culpably* attack B or *responsibly* attack B, since the leaders of A are subject to mind control and are being directed to do things that they otherwise would never do. If we accept the view that it's permissible for B to attack A if and only if A has committed a serious wrong against B, we'll have to say that it's permissible for B to attack A in this case. But, if we stick to the analogy with individual self-defense, we might see some reasons for balking at this. After all, last week we saw an extended defense by Otsuka on which it is not permissible (though fully excusable) to attack a nonresponsible, nonculpable threat in self-defense. Given the domestic analogy, we can generalize this conclusion: if so, we must agree that the initial theory at issue was too permissive.

2.2. *The Requirement of Competent Authority*

Principle II of orthodox *jus ad bellum* holds that a war may be initiated only by those who are authorized to do so. This principle is a seriously questionable, since it is easy to argue that it's either obviously false or trivially uninteresting.

If "authorized" in the principle means "authorized by law" or something else essentially *conventional*, then the competent authority requirement rules out the permissibility of acts of war that are parts of a revolutionary overthrowing of a corrupt or tyrannical government, since the initiators of these acts will *not* be conventionally authorized to act in the way they're acting. So, if the principle is to be plausible, "authorized" has to be understood in some other, nonconventional sense. But what could that sense be? The only sense that seems sufficiently nonconventional to avoid this first problem is "morally authorized". But if "authorized" means *this*, then the competent authority requirement just collapses into the principle of Just Cause, since anyone who has an all-things-considered just cause for war will trivially be morally authorized to act on it. Either way, the Principle II doesn't seem to be a serious *extra* constraint.

Notably, this is not to say anything about how we ought to design the law. If it were easy to arrange a war without a fairly drawn-out process of conventional authorization, there would reasonably be worries about abuse and the legal permission of wars that aren't actually just. So, to say that Principle II is not fundamentally morally significant in particular cases isn't to say that it shouldn't be highly significant to how it would be morally best to set up the law. (This parallels our discussion of the morality vs. the right law of voluntary euthanasia.)

2.3. *The Requirement of Right Intention*

Principle III of orthodox *jus ad bellum* holds that a war may be initiated only by those who intend the war to occur for the *right reasons* -- e.g., that it would be just. This needs to be qualified a little bit. It is too strong to demand that a war be intended *only* for the right reasons. After all, as Jeff points out, it is permissible for you to stop a mugging if you want to get a reward *and also* want the victim to be better off for his own sake. So, by analogy, one might think that at least some amount of mixed motivation is acceptable in the case of planning a war, and some subset of the reasons needn't be ones that would *by themselves* justify the war. Of course, here we do want to distinguish morally *bad* motives from morally *irrelevant* ones; mixed motivation that includes bad motives may be less obviously OK.

All the same, this requirement is not strictly correct even when it's qualified. As I've stressed in many classes, permissibility and praiseworthiness come apart, as do impermissibility and blameworthiness, and intentions are only relevant to praise and blame. If there really are decisive reasons for a state to go to war, it does the right thing in doing so. The people who initiate the war may be less praiseworthy or even quite blameworthy if they don't care about the justness of the war. But if there's a just cause, war ought to be permissible. If we happen to have leaders who don't ultimately care about the good and the right, and our state is wrongfully attacked, the fact that the leaders won't initiate a response except for bad reasons shouldn't lead us to conclude that the only permissible thing to do is not engage in a defensive attack at all. Of course, we can grant that it would be *better* if they would (or could, in the relevant case) have the right intentions: the point is simply that their intransigent moral myopia should not render *impermissible* a defensive attack when it really would be justified from an impartial point of view.

After all, the same goes for individual self-defense. If a murderer prepares to attack you and you get ready to fight back, but not mainly out of self-interest but rather out of hatred (which, we can stipulate, is the only factor that is psychologically capable of motivating you), that shouldn't make it impermissible for you do anything except let yourself be slaughtered.

2.4. *The Requirement of Necessity*

Principle IV of orthodox *jus ad bellum* holds that it is not morally permissible to resort to war if it isn't necessary. This principle is more plausible and crucial than the last two, though it requires some interpretive commentary. For, at least if "necessary" is understood in one natural sense, the principle is false. I take it that it would be natural to say that if there are some *possible* means M alternative to M* of achieving some end E, then M* isn't a *necessary* means for E. But there are many *possible* means of achieving some end that wouldn't be *effective* means. And when the probability of successfully achieving E given M* is much less than the probability of achieving E given M, it is plausible that M ought to be preferred as a means to M*.

So, "necessary" here really means "necessary insofar as we're looking only at the set of effective means". In his review piece, Jeff raises a further objection to the principle, but my intuitions about the case aren't as clear as his:

> Suppose that a state has a just cause for war…that it could pursue this just cause either by war or by non-belligerent means, and that both means have the same probability of being equally effective. Suppose, however, that the peaceful means would require even greater sacrifices than war (for example, great economic costs), while war promises certain compensations (for example, the ability to force the aggressor to pay appropriate reparations). Suppose that only this state has the power to achieve the just cause and that it would be better, from an impartial point of view, for the state to fight the war than to allow the just cause to go unachieved. But suppose, finally, that the state is not morally required to pursue either means of achieving the just cause…. The requirement of necessity says that it is not permissible for the state to go to war; yet that seems wrong.[32]

[32] "Just War", pp.673-674.

Thoughts? While I share Jeff's gut reaction to some degree, I think we have to be careful in rejecting theories on the basis of intuitive responses like this.

As we saw in the discussion of maximizing consequentialism, the view entails that if some act A would bring about just *slightly* better consequences than some other act B (say, 100 units of goodness as opposed to 99.9), it's impermissible to do B. Some people think this is a sufficient reason for rejecting the view. But, as I noted, maximizing consequentialists have many resources at their disposal to explain away the gut feeling of implausibility. One thing they can do is note that although they, like virtually everyone, accept an all-or-nothing understanding of the permissibility/impermissibility divide (so that, for every act A, A is either definitely permissible or definitely impermissible), they could easily accept, like everyone else, a graded picture of the rightness/wrongness divide. If they did, they could say that if A brings about better consequences than B, then B-ing would be less right than A-ing in proportion to the degree to which A's consequences are better than B's. They could then say that although B-ing in this case is strictly speaking impermissible (since impermissibility is an all-or-nothing matter), this doesn't imply that B-ing is wrong to any *significant degree*; it's just that there is a slightly better option, and that's the only option one ought (and hence is permitted) to pick. Moreover, if B-ing wouldn't bring about any *bad* consequences but would only bring about *less good* consequences than A-ing, maximizing consequentialists can grant that an agent would be at worst only negligibly blameworthy for picking B over A, and at best not blameworthy at all, but just less praiseworthy. As I was suggesting in the earlier class, once these concessive points are borne in mind, the maximizing consequentialist's verdict about the original case seems less crazy.

Precisely the same kind of point can be made about Jeff's case. Choosing the peaceable option would be *slightly* impartially better than going to war, and so, if we treat impermissibility as an all-or-nothing matter, the necessity requirement entails that going to war is all-things-considered impermissible, given that, by stipulation, both options are equally effective for achieving the just cause. But this isn't to say that the state in his example would be blameworthy to any noteworthy degree for going to war, or that it would be *seriously wrong* to do so: it would be just a bit less right, and so a bit less praiseworthy, than the peaceable option. Once these concessive points are borne in mind, the necessity requirement's verdict about Jeff's case seems less silly.

2.5. *The Requirement of Proportionality*

Principle V of orthodox *jus ad bellum* holds that the relevant expected good effects that a war can be expected to achieve must be sufficiently important to justify causing the relevant expected bad effects. Like the requirement of necessity, this principle is much more plausible and crucial than II and III, though it, too, requires some interpretive commentary.

The key word that needs to be clarified in the principle is "relevant". What good effects and what bad effects count as "relevant"? Without an answer to this question, there is no clear way to apply this principle.

With respect to the *good* expected effects, "relevant" means "pertinent to the achievement of the just cause". Good effects that are irrelevant to the achievement of the just cause should probably not figure into calculations of proportionality. As Jeff nicely puts it: "That a war

would teach valuable lessons about comradeship, what matters in life and so on to some of the soldiers, or give them exhilarating experiences of combat, is not a consideration that can weigh against and help to justify the incidental or unintended killing of innocent civilians on the other side".[33]

With respect to the *bad* expected effects, "relevant" seems harder to define in any systematic way. Most bad effects will probably be relevant, but the *degrees* of their relevance will be highly variable. For instance, if we know that civilian casualties will reach a certain high number N, that is a factor that is relevant in a more weighty fashion than the fact that deaths of unjust combatants will reach N. As a rule, perhaps it would be right to say that if certain bad effects are *not essential to achieving the just cause in the sense that, by bringing about those effects, the cause is directly advanced*, those effects should have greater weight in proportionality calculations. So, for instance, though we may not be able to avoid stray civilian casualties, bringing about these casualties does not *directly* advance the just cause, but is just an indirect, contingent side-effect of it. That's why these casualties are weightier in calculations than the deaths of unjust combatants.

2.6. *The "Reasonable Hope of Success" Requirement*

Principle VI of orthodox *jus ad bellum* holds that there must be a reasonable hope of succeeding in pursuing the just cause. As Jeff has pointed out, this principle is redundant once the requirement of necessity is properly understood. For, as he noted, in thinking about whether war is necessary for achieving a just end, we restrict our focus to the *effective* means to it insofar as there are any. Once our focus has been restricted in this way, we automatically end up claiming that the resort to war is permissible only if there is a reasonable hope of achieving the just cause, since, by assumption, the resort to war has to be an effective means of achieving that cause.

3. ***Jus in Bello*** **and the Traditional/Revisionist Divide**

Let's turn to the principles of *jus in bello* with an eye to the differences between the Traditional and the Revisionist Theories. Defenders of both theories agree that there are at least three central principles of *jus in bello*: the requirement of *discrimination*, the requirement of *proportionality*, and the requirement of *minimal force*.

Stated very generically (and quite unhelpfully!), the first requirement is:

> *Discrimination*: It is only permissible to attack legitimate targets.

What targets are *legitimate*? This is an issue about which defenders of the Traditional and Revisionist Theories disagree. Defenders of the Traditional Theory uphold:

> *Traditional Target Legitimacy:* A target is legitimate if and only if it is not innocent.

What is an *innocent* target? Here, "innocent" is being used as a technical term derived from the substantive plural participle of the Latin verb *nocere* – i.e., *nocentes*, which means "the harming ones". An innocent target in this technical sense is a target that doesn't fall among

---

[33] Ibid., p.674.

the *nocentes*: so, in brief, it's a target that isn't engaging in any harming. Friends of the Traditional Theory thus often identify the innocent with *noncombatants*. Understood in this fashion, Discrimination and Traditional Target Legitimacy entail the following principle:

> *Noncombatant Immunity*: It is not permissible to attack noncombatants.

Noncombatant Immunity is pretty plausible, though just a bit shaky. Workers in a munitions factory are not directly harming anyone. But they are indirectly contributing to harm, and to the war effort. So, it is a little odd to say that they're fully innocent, at least when they're at work in the factory. Some might think it's obvious that a just combatant would be permitted to blow up a munitions factory if that would clearly advance the right cause. But if Noncombatant Immunity were true, this would be false. This doesn't show that Discrimination and Traditional Target Immunity are false: it just shows that defenders of the Traditional Theory may need to revise their understanding of innocence to prevent the derivation of Noncombatant Immunity.

Revisionists reject Traditional Target Legitimacy. Why? The core reason is that they deny that just combatants lose their right not to be killed by engaging in defensive attacks, just as any sensible view of individual self-defense would deny that someone defending himself from a liable attacker loses his right not to be killed. Hence, defenders of the Revisionist Theory claim that just combatants are not legitimate targets for attack by unjust combatants, and hence deny:

> *The Moral Equality of Combatants*: Whether a combatant is on the just or the unjust side in a war is irrelevant to whether his individual acts of war are (im)permissible.

While the denial of this claim can seem a little surprising at first, the combination of the "domestic analogy" with the point that a war *simply consists* of individual acts of war makes it pretty appealing, and makes Traditional Target Legitimacy look far less obvious.

Of course, much more will need to be said to defend this view, and we'll get to that in a moment. The other thing worth noting, though, is that whether one accepts or rejects the *Moral Equality of Combatants* has significant implications for how one understands the other key element of *jus in bello*, which is the requirement of proportionality.

According to this requirement, an individual act in a war is permissible only if the relevant expected good effects that the act can be expected to achieve are sufficiently important to justify the relevant expected bad effects. If "good" and "bad" are here understood objectively rather than subjectively, it becomes strikingly hard to see how combatants on the unjust side could ever avoid violating the requirement of proportionality *if* we reject the Moral Equality of Combatants. After all, it will be clearly objectively bad for a just combatant to be killed, and the expected effects of this couldn't be objectively good, since they could only help to advance the unjust cause. So, even setting aside the requirement of discrimination, the requirement of proportionality will entail that most acts of war by unjust combatants are morally impermissible.

If, on the other hand, one accepts the Moral Equality of Combatants, this reasoning will not obviously go though. For, given that it will now be permissible for an unjust combatant to

kill his just attacker, the costs of killing that attacker will not fall out as being as bad as they would have been according to the Revisionist Theory. Still, given that the unjust combatant still lacks an ultimately good aim, it remains hard to see how he could ever avoid violating the requirement of proportionality. This, I think, is what's even more puzzling about the Traditional Theory's claim that *jus ad bellum* and *jus in bello* are fully independent: if a war is unjust and hence fails to abide by the principles of *jus ad bellum*, no individual acts that contribute to the unjust side's winning the war could be objectively good, and so it will be impossible for any given unjust combatant to abide by all the principles of *jus in bello* if they are understood objectively.

The same sorts of points hold for the last standard requirement of *jus in bello* – i.e., the rule of minimal force. This requirement is, in effect, the *in bello* analogue of the *ad bellum* requirement of necessity, and states that an individual act of war is permissible only if the harms that it causes are necessary to do enough good in advancing the cause of war to outweigh the badness of these harms. Obviously, if a state's cause is not just, it doesn't make sense to say that the combatants who seek to advance that cause would do *any* relevant good in harming for the sake of the cause.

4. The Case for Revisionism

There are many arguments for the Revisionist Theory, and I'll focus on just one of them today – one we've already seen to some degree. It is really an argument against the Independence Thesis, which is an essential component of the Traditional Theory:

Against the Independence Thesis, Part I

1. As understood by the Traditional Theory, the Independence Thesis implies that individual acts of war by combatants on the unjust side in a war can be entirely morally permissible.

2. But the individual acts of war by the combatants on the unjust side in a war cannot be entirely morally permissible. Indeed, they are *of necessity* often morally impermissible.

    a. Argument for (2):

        i. If the "domestic analogy" holds (i.e., the analogy between justified individual self-defense and justified state-defense), then the state on the unjust side of a war cannot permissibly attack the state on the just side when that side takes defensive action.

        ii. It is incoherent to claim that the state on the unjust side cannot permissibly attack the state on the just side *but* that the individual acts of war by combatants on the unjust side can be often morally permissible. After all, these individual acts of war collectively *constitute* the attack the unjust side is making against the just side.

        iii. (2) follows from (i) and (ii).

3. So, the Independence Thesis, at least as understood by the Traditional Theory, is false.

There are many replies that Traditional Theorists give to this argument. One thing the esteemed traditionalist Walzer has said in objecting to premise (ii) in the sub-argument for (2) is that, while leaders of a state have a choice about starting a war, prospective combatants do not have much of a choice about participating in it. As he says: "Personal choice effectively disappears as soon as fighting becomes a legal obligation and a patriotic duty…. For the state decrees that an army of a certain size be raised, and it sets out to find the necessary men, using all the techniques of coercion and persuasion at its disposal."

In reply to this reply, I agree with and simply defer to Jeff: "[T]hese considerations are beside the point. For the various considerations Walzer cites are at best *excuses*. They may show that a particular unjust combatant is not a criminal and is not to be blamed or punished for what he does, but they do not show that he acts permissibly. If…unjust combatants are at best merely excused for fighting, while just combatants are justified, two of the central tenets of the traditional just war theory must be rejected. It is false that unjust combatants do no wrong to fight provided they respect the rules of engagement. And it is false, *a fortiori*, that *jus in bello* is independent of *jus ad bellum*."[34]

On closer inspection, *most* of the arguments that friends of the Traditional Theory give against (ii) in the argument for (2) rest on obvious conflations of justification and excuse. Another factor that people routinely appeal to is the limited information that many combatants possess: "It is often noted that soldiers cannot have access to all the information and arguments relevant to the justification for war, and even if they could, they could not be held responsible for reasoning successfully about subtle and contested matters of *jus ad bellum*; therefore it is reasonable for them to defer to the judgment of their leaders and permissible for them to fight."[35] As we've noted many times, the fact that an agent has misleading information is relevant only to his blameworthiness for acting, not to the permissibility of his acts. While nonculpable ignorance is an excuse, it isn't a justification.

Another reply that is sometimes advanced against (ii) in the argument for (2) does not so obviously conflate justification with excuse. This reply begins with the reasonable observation that the military is an institution that does have value and ought to be sustained. After all, if a state *is* confronted with a defensive attack, and all the criteria for *jus ad bellum* are satisfied, it ought to fight back, and it will need a functioning military to do that. But, defenders of the reply point out, "if the institution is to survive and carry out its functions, others within it must fulfill their assigned roles even if they disagree with the decisions reached by those responsible for matters of *jus ad bellum*". So, the reply goes, there is a genuine obligation grounded in the overall value of sustaining the military institution for combatants to heed the orders they receive.

The main problem with this reply is that it just isn't plausible that this institutional obligation will be strong enough in the case in which combatants really are part of an unjust war to outweigh their standing moral obligations not to kill those who haven't lost their right not to be killed, and the just combatants will be among these people. As Jeff puts it: "[I]nstitutional

---

[34] McMahan (2004: 700).
[35] Ibid., p.703.

obligations are insufficiently weighty to override the duty not to kill…as a means of achieving unjust aims."[36]

A simpler counter-reply is that institutional obligations are *cancelled* when the institution adopts an unjust aim. This is vivid in the case of Nazi Germany. As Jeff puts it, generalizing the point to less extreme cases: "The Nazi military, for example, was incapable of imposing moral duties on those who occupied roles within it. On occasion the orders that a member of the Nazi military received coincided with what that person was morally required to do, but the source of that requirement was never a duty to maintain the functioning of the Nazi military. The same is true of all other military institutions…that are designed and structured for external aggression and internal repression, torture and murder. So, even if most soldiers have duties to fulfill the functions of their role even when they are commanded to fight in an unjust war, not all do."[37]

5. Limits to the Revisionist's Denial of the Moral Equality of Combatants

While we'll continue to discuss further arguments for and challenges to the Revisionist Theory's stance on the Moral Equality of Combatants next week, it is important to realize right away that there are some built-in limits to the Revisionist Theory's stance. The Revisionist Theory only denies the Traditional Theory's claim that just and unjust combatants are ***always*** morally on a par, and is not committed to denying that there are some special cases where they may be on a par.

Of course, in *some* types of example, the Revisionist Theory will make a much bolder claim. In virtually every case of defensive action against an unprovoked, unjust attack, the Revisionist Theory will indeed deny that any unjust combatants have undefeated rights not to be killed, while the just combatants retain their undefeated rights not to be killed. But not all just wars involve defensive action. One can imagine a case in which one state A unjustly seized land from another state B, and enough time elapsed that the people who were originally responsible for the unjust seizure have died of natural causes. Nevertheless, if B continues to possess this land without any *ex post facto* authorization from A, one might think that A would be permitted to engage in a militaristic attempt to regain the land of which it is really the rightful owner. But given that none of the people who were directly responsible for the unjust seizure are living anymore, it seems odd to say that the current inhabitants of the unjustly seized land have done anything to lose their rights not to be harmed as a necessary and proportionate component of A's justly regaining its land. If that's right, it seems a little harder to see on what grounds the Revisionist Theory could claim that the current inhabitants of the land who will serve as fighters for B are not at least *close* to being morally on par with the just combatants from A.

So, it seems to me that there may well be types of conflicts for which even the Revisionist Theory shouldn't reject a restricted version of the Moral Equality of Combatants. And, given the Revisionist Theory's reasons for rejecting the unrestricted version of this thesis, this concession does not strike me as at all problematic or incoherent.

---

[36] Ibid., p.708.
[37] McMahan (2009: 73).

1. A Quick Recap: The Revisionist Critique and Some Traditionalist Responses

Today we're going to continue to examine some responses that defenders of the Traditional Theory of the just war have offered to the foundational criticisms mounted by Revisionists like Jeff. First, though, let me quickly recap the core of the Revisionist Theory's onslaught, and a couple of the responses that we started to examine last week.

As you'll recall, the main claim of the Traditional Theory that Revisionists tend to attack is:

> *The Moral Equality of Combatants* (M=C): Whether a combatant is on the just or the unjust side in a war is irrelevant to whether his individual acts of war are morally permissible or impermissible.

Revisionists' strongest argument against M=C goes like this:

1. If M=C is true, then individual acts of war by combatants on the unjust side in a war can be entirely morally permissible.
2. But if the "domestic analogy" holds (i.e., the analogy between justified individual self-defense and justified state-defense), then the state on the unjust side of a war cannot permissibly attack the state on the just side when that side takes defensive action. More generally, the state on the unjust side of a war can't permissibly attack the state on the just side in the war when the state on the just side rightly responds to the wrong that the unjust side has committed.[38]
3. Yet it is incoherent to claim that the state on the unjust side cannot permissibly attack the state on the just side *but* that the individual acts of war by combatants on the unjust side can be often morally permissible. After all, these acts of war collectively *constitute* the attack the unjust side is making against the just side!
4. So, given (2), (3) and the domestic analogy, it follows that there are at least some (indeed, many) cases in which the acts of the combatants on the unjust side in a war are not as permissible as the acts of the combatants on the just side. And that's just to say that M=C is false.

The main response that Traditionalists give to this argument is to question whether the assertion that Revisionists claim to be incoherent in (3) really is incoherent.

We saw two types of questioning by Traditionalists last week. One type seems to confuse justification with excuse. Traditionalists note that combatants on the unjust side are often

[38] Revisionists do have to be a little careful about moving from the first to the second sentence in (2). For, as we saw at the end of last week's meeting, some just wars that do not involve defensive action seem to pose exceptions. In the case I used last time, one state A unjustly seized land from another state B, and enough time elapsed that the people who were originally responsible for the unjust seizure have died of natural causes. As I suggested, if B continues to possess this land without any *ex post facto* authorization from A, one might think that A would be permitted to engage in a militaristic attempt to regain the land of which it is really the rightful owner. But given that none of the people who were directly responsible for the unjust seizure are living anymore, it seems odd to say that the current inhabitants of the unjustly seized land have done anything to lose their rights not to be harmed as a necessary and proportionate component of A's justly regaining its land. If that's right, the second sentence in (2) is false and needs to be restricted.

unaware that their side really *is* the unjust side and believe that they're on the just side. Alas, while it's uncontroversial to say that nonculpable ignorance of a wrong is an excuse, it has historically been controversial to say that it's a justification. The same goes for Traditionalists' appeal to the fact that many combatants on the unjust side are fighting under duress, and face the choice between fighting for their country or being executed or locked up. That's a great excuse, but it isn't a justification – it doesn't make it OK for them to kill unjustly, but just relieves them of blame for doing so, since they're really the coerced instruments of the state.

The other type rests on an appeal to institutional obligations. If it were easy for soldiers to refuse to go to war and to question and undermine the decisions of their leaders, it would be difficult to sustain a functioning military. Since there seems to be value in having a functioning military (after all, one might get attacked by Nazis), there is a general reason for soldiers to give into orders and not to try to question and undermine the decisions of their leaders. This reason, according to Traditionalists, provides a kind of rule consequentialist justification for the acts of combatants in unjust wars. The main problem with this reply is that this reason is just too weak not to be massively outweighed in particular cases: institutional obligations are not going to trump the moral reasons that Nazi combatants had not to commit their individual acts of war.

Today I want to talk at first about a couple of further, slight better types of response to the Revisionist critique. One of them is essentially a more sophisticated, conflation-free version of the first one I already discussed. Some people really do think that 'ought' claims – even moral 'ought' claims – are evidence-relative in such a way that the fact that someone falsely but rationally believes that his war is just entails that he ought to act in it. This theory is becoming increasingly popular (though it is not yet a majority view), and some people think that it does not merely amount to a conflation of justification and excuse. If they're right, then Jeff's and my own response to the appeal to unjust combatant's nonculpable ignorance of the wrongness of his cause as a justification for his acts doesn't really work. The second new response I'll discuss appeals to the idea that combatants who have signed up for the military and who deliberately put on uniforms actually give some kind of *consent* to be regarded as legitimate targets. Since we have seen in discussions of other topics (e.g., voluntary euthanasia) that a person's consent to be killed can partly explain why it is permissible to kill him, some defenders of the Traditional Theory think that it can help to explain the Moral Equality of Combatants.

2. Evidence, 'Ought' Claims and the Charge of Permission/Excuse Conflation

In recent years, some have claimed that whether some sentence of the form 'Person S ought to do A' is true sometimes depends on S's evidence. These people are very self-conscious about not simply conflating permission and excuse. Indeed, they think their view is supported by the folk's semantic intuitions about 'ought' claims. To see this, consider the following famous case that I take *verbatim* from a classic paper by Frank Jackson:

> *The Drug Example*. Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin ailment. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the literature has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it. One of the drugs B or C will completely cure the

> skin condition; the other will kill the patient, and there is no way that Jill can tell which of the two is the perfect cure and which the killer drug.[39]

What ought Jill to do? On the face of it, the overwhelmingly intuitive thing to say is that she ought to give John drug A. Notably, however, drug A is not only *not* the best cure, but it's actually *known* not to be the best cure – though which cure really *is* best isn't known either. Jackson and other philosophers take this to establish that some 'ought' claims are not determined just by the agent's circumstances, but also by the limited evidence he possesses.[40] This, they think, also suffices to undermine the idea that the only properties we need to formulate a complete normative theory are objective obligatoriness and blameworthiness and praiseworthiness. After all, the intuition here isn't that Jill would merely not be blameworthy for choosing A. It's straightforwardly that she *ought* to choose A.

Importantly, it doesn't follow just from these intuitive arguments from cases like *The Drug Example* that *no* 'ought' claims are determined by the agent's external circumstances. Indeed, there is considerable pressure to grant that some 'ought' claims are not evidence-relative at all even if some are. To bring this out, consider the following further case, which I take *verbatim* from a recent paper by Niko Kolodny and John MacFarlane:

> *Mine Shafts.* Ten miners are trapped either in shaft A or in shaft B, but we do not know which. Flood waters threaten to fill the shafts. We have enough sandbags to block one shaft, but not both. If we block one shaft, all the water will go into the other shaft, killing any miners inside it. If we block neither shaft, both shafts will fill halfway with water and just one miner, the lowest in the shaft, will be killed.[41]

As Kolodny and MacFarlane note, it seems that the outcome of our deliberation should be:

i. We ought to block neither shaft.

This is much like in *The Drug Example*, since this is known not to be the best option, but, given our ignorance, it is the least risky and hence, it seems, the obligatory option. Strikingly, though, Kolodny and MacFarlane compellingly suggest that the following claims are equally plausible things to accept in deliberating about what to do in this case:

ii. If the miners are in shaft A, we ought to block shaft A.

iii. If the miners are in shaft B, we ought to block shaft B.

Since the miners are either in shaft A or shaft B, it logically follows from (ii) and (iii) that:

iv. We either ought to block shaft A or we ought to block shaft B.

But (iv) is directly inconsistent with (i). So, it looks like we've got a paradox. A natural idea that some people (though not Kolodny and MacFarlane themselves) accept is that there is just an equivocation here: the 'ought' in (i) is evidence-relative, while the 'ought' in (ii), (iii)

---

[39] Jackson (1991: 462 – 463).
[40] See also Zimmerman (2008) for a lengthy defense of the evidence-relative view of 'ought' claims.
[41] Kolodny and MacFarlane (ms: 1-2).

and (iv) is objective. If 'ought' were ambiguous or context-sensitive like this, there would be no inconsistency between (i) and (iv), since 'ought' wouldn't mean the same thing in each. As we'll see shortly, as long as we accept that there is both an objective sense of 'ought' and an evidence-relative sense of 'ought', we can raise problems for appeals to the latter sense to justify the Traditional Theory's commitment to the Moral Equality of Combatants.

But first let's see how the evidence-relative 'ought' might seem to help the Traditional Theory. If we accept that there are some true 'ought' claims that are made true not just by a subject's circumstances but by facts about his limited evidence and irremediable ignorance, we ought to think, when a combatant X on the unjust side has limited access to information about the nature of the conflict in which he is fighting, and this information in fact suggests that he is on the just side, that there are some true claims of the form 'X ought to kill Y', where Y is a just combatant. But if it's true that X ought to do this, then it trivially follows that X is permitted to do it. Accordingly, defenders of the Traditional Theory might reason, we get indirect evidence for (an at least restricted) version of the Moral Equality of Combatants, since informational limitations will make it the case that some soldiers on the objectively unjust side are in fact morally permitted to attack soldiers on the just side (and *vice versa*, and hence the equality).

This, I think, is a pretty interesting point. It's important to note, however, that it really is clearly limited in ways that Traditionalists have to admit, and falls short of realizing the full ambitions of their view. While *some* unjust combatants may lack access to evidence that establishes the unjustness of their cause, this is certainly not always the case. As Jeff puts it, referring to the appeal to the evidence-relative 'ought' as the "epistemic argument":

> Even the epistemic argument for a modest version of [M=C] fails. For it to work, it would have to be true in *every* unjust war that it is…reasonable for most combatants to believe that their war is just. Only that way would most of them be subjectively justified in fighting, thus establishing a presumption that *each* acts with subjective justification in conditions in which it is impossible to determine, of any given unjust combatant, whether his beliefs are in fact…justified. But there are some wars that are so obviously unjust that it is simply not possible that the majority of those who fight in them are justified in believing that they are just. A majority may, of course, *believe* that such a war is just, but that does not mean that they have…*justification* [i.e., *evidence*] for believing it. It may be, for example, that a majority of the Nazi soldiers who invaded Poland, or Czechoslovakia, or France, or any of the various other countries they invaded, believed that their campaigns of aggression and extermination were just, but it would be absurd to suppose that in most cases those beliefs were reasonable in the context.[42]

And, crucially, cases like *Mine Shafts* and *The Drug Example* establish only that 'ought' is evidence-relative, not that it's merely belief-relative. So, even if a fully objectivist analysis of 'ought' were not viable, this couldn't help the cause of the Traditional Theory. Of course, not all cases will be as extreme as the case of Nazi Germany, but it only takes some cases to refute the Moral Equality of Combatants, which is a fully universal claim.

---

[42] McMahan (2009: 65).

There are larger problems with the appeal to the evidence-relative 'ought' in defense of the Traditional Theory. The evidence-relative 'ought', assuming there is one, does not interact with rights in the way that the Traditionalist needs it to. To bring this out, consider the case of individual self-defense against a nonculpable and misled threat. Suppose that a mad scientist has hijacked the visual processing centers of Bill's brain, and makes him hallucinate that I am a huge bear about to rip him to shreds, when I am in fact just about to bring him a box of chocolates. If the evidence-relative view is true, it's true that Bill is permitted to attack me. After all, his evidence decisively suggests that I'm going to kill him, and he has no way of knowing otherwise. Nevertheless, it *does not* follow from these facts that I have no right to defend myself from him when he whips out his revolver and prepares to blow my brains out, and that, more generally, no one else has a right to prevent him from attacking me. A third party would be permitted to intervene to stop him. Bill is still a responsible agent, since the only thing that's gone wrong with him is his visual processing, and he's otherwise rational and autonomous and so on; it's not like he's under the influence of an autonomy-reducing drug. Notably, though, if a third party were only able to interfere with my defensive action by injuring me, they would not have a right to do so here. So, there is a conspicuous moral asymmetry between me and Bill *in spite of the fact (if it is a fact) that he is permitted in the evidence-relative sense to attack me.*

And this gets close to letting us reinstate the Revisionist's critique of M=C. As Jeff puts it, introducing some slightly different terminology but making basically the same point:

> It is common to distinguish between two types of right: liberty rights and claim rights. A liberty right is merely a permission. To say that a person has a right to do *x*, when what is meant is that she has a liberty right, is just to say that it is not wrong for her to do *x*. But a claim right is not [just] a permission but a right against intervention. To say that a person has a claim right to do *x* is to say that no one else has a liberty right to prevent her from doing *x*.… *The claim that unjust combatants are subjectively justified in fighting against and killing just combatants is at most an assertion of a liberty right. [Nonculpable ignorance] may ground a permission to act, but it cannot possibly ground a right against interference. By contrast, just combatants do have a claim right to fight by permissible means in support of their cause.*[43]

So, there remains a major moral asymmetry between just and unjust combatants, and it follows from this that at least part of the Traditional Theory has to be false *even if* it's true that 'ought' claims are evidence-relative. The rules of *jus in bello* will not be independent of the rules of *jus ad bellum*, and it will not be true that just and unjust combatants are morally on a par when it comes to their individual acts, even if the unjust combatants have misleading evidence.

Finally, to put the nail in the coffin, recall that what Kolodny and MacFarlane's reasoning about the *Mine Shafts* case plausibly shows is that there have to be two senses of 'ought', and hence that 'ought' is not always evidence-relative. If that's right, then even in the cases where unjust combatants are permitted in the evidence-relative sense to attack just combatants, it will remain an open question whether they are permitted in the objective, non-evidence-relative sense to attack just combatants. If not (and the analogy with *Mine Shafts* suggests not), the Moral Equality of Combatants will still fail in the sense that there

[43] Ibid., p.63.

will be a perfectly eligible alternative sense in which it just isn't true that unjust combatants may kill just combatants.

3. Consent, the 'Boxing Match' Model and the Moral Equality of Combatants

Let's turn to a further and completely different attempt to rescue the Traditional Theory's endorsement of M=C. This attempt rests on the thought that by entering the military and putting on uniforms that explicitly distinguish themselves from noncombatants, just combatants consent at least some degree to the possibility of being killed, even if unjustly. Jeff nicely refers to this idea as the 'boxing match' model of combat in war: "[On this view], war is analogous to a boxing match or a duel. Just as it is part of the profession of boxing to consent to be hit by one's opponents, so it is part of the profession of arms to consent to be attacked by one's adversaries."[44] Less crudely, Jeff states (though clearly doesn't endorse) this idea as follows:

> [G]overnments demand that their soldiers commit themselves to fight in any war they may be ordered to participate in. But if in voluntarily joining the military, soldiers are agreeing to fight in any war, just or unjust, they must be accepting a neutral conception of their role, according to which they are permitted to kill their adversaries, irrespective of whether the latter are just or unjust combatants. And elementary consistency requires that they recognize that all other combatants are in the same situation and have the same privilege. They concede, in other words, that their adversaries are permitted to kill them whether they happen, on any particular occasion, to be just or unjust combatants.[45]

There's another component to this idea at which I already hinted, which Jeff summarizes:

> There is…a further reason to suppose that soldiers in fact waive their right not to be killed. Combatant status has conventional and legal dimensions, the most important of which is the requirement that combatants openly identify themselves as such, usually through the wearing of a uniform. When a soldier puts on the uniform, he [seems to be] consciously identifying himself as a legitimate target…. But [it may seem that] to identify oneself conspicuously as a legitimate target of attack just *is* to consent to be attacked.[46]

Now, if all this is right, it might seem like there ought to be an argument that the Moral Equality of Combatants is a *de facto* truth even if there are possible worlds in which it's *de facto* false. Here, I take it, is the argument (the flaws of which are clearer when it's put formally):

> I. By signing up for military service and wearing uniforms in combat, combatants consent to be legitimate targets in any war, just or unjust.
>
> II. If combatants on one side consent to be legitimate targets in any war, just or unjust, they make it permissible for combatants on the other side to kill them, even if these latter combatants are unjust.
>
> III. If (II), then the Moral Equality of Combatants is a *de facto* truth.
>
> IV. So, the Moral Equality of Combatants is a *de facto* truth.

---

[44] Ibid., p.52.
[45] Ibid., p.53.
[46] Ibid., p.55.

There are a lot of problems with this argument. But it's important to realize that even if we grant the core idea that there is consent, there is no reason to accept premise II.

While it's true that consent is often a necessary condition for permissible killing (as in the case of euthanasia, if tradition is right that involuntary euthanasia is impermissible), and it's true that a person's consenting to be killed can often play a role in explaining why it's permissible to kill him, it does *not* follow that consenting to be killed is a sufficient condition for making yourself a permissible target of killing.

There are two reasons for this. One of them is that, when X's consent to be a victim of some act is a product of coercion, duress, or ignorance, it plays no role in making it permissible for that person to be victimized accordingly. Consider two cases to appreciate this point. Firstly, suppose that a serial killer and rapist decides that he is going to let his victim have a choice about which of the two wrongs he'll perpetrate: if she doesn't consent to being his sex object for the next week, he will kill her, but if she does consent, he'll leave her alone after a week. If she consents, that really doesn't make it morally OK for him to use her as a sex object for the next week. So, consent to being a victim is clearly not sufficient to make it permissible for you to be accordingly victimized when it's a product of coercion or duress.

Now consider a second case. Suppose an otherwise reliable acquaintance has a change of heart and suddenly wants, unbeknownst to you, for you to die. He comes to you in the night and tells you, falsely, that you are about to be kidnapped by terrorists, tortured for a week, and then burned to death. He does note that he could give you a lethal injection and prevent you from being harmed and killed in such a gruesome way. You believe him, and consent to the lethal injection. It seems clear that this doesn't make it suddenly morally OK for him to kill you. So, when consent to being a victim is a product of ignorance, it needn't be sufficient to make it permissible for you to be accordingly victimized.

But, notably, a combatant's consent to being a legitimate target (assuming he really does give it by joining the military on some occasion) may, in some cases, be sustained only by coercion, duress or ignorance of the normative or nonnormative facts. So, when a soldier recognizes that his opposition is unjust but stays in the military only because he feels compelled by patriotism (or by the costs of trying to get himself out of the military, or because he does not realize exactly how unjust his opposition really is) to fight, it needn't be true that the consent he initially offered by joining the military (assuming again that there was such consent) is sufficient to make it permissible for him to be killed by combatants on the other side.

There is a deeper reason why consent isn't sufficient for permissible killing. Jeff brings it out nicely in his book, and I must simply defer to him on the point:

> Suppose…that it is true both that just combatants consent to be killed and that this means that unjust combatants who kill them neither wrong them nor violate their rights. It still does not follow that unjust combatants act permissibly when they kill just combatants. That acts of war by unjust combatants kill or injure just combatants is not the only reason they are morally objectionable. These acts are also instrumental to the achievement of an unjust cause and thus wrongfully

> threaten people other than the just combatants who are the immediate targets of attack…. So even if we assume that all combatants consent to be killed…there is still no equality of moral status between just and unjust combatants, since the military action of unjust combatants supports an unjust cause and threatens innocent bystanders with harm that is not outweighed by the good effects that their action might be expected to achieve.[47]

This provides a decisive reason to reject premise II in the argument considered above.

That is not, however, the only problem with the argument. Claim I also seems problematic for at least three reasons. The first is that there seems to be no support for I once we recognize a distinction between *acceptance* and *consent*. By putting on a uniform, a combatant does accept in a purely conventional sense that there is a difference between him and a noncombatant, and that, if it's a choice between having him be attacked and having a noncombatant be attacked, the former ought to be what occurs. But to accept these facts is not to provide someone else with the permission to kill you, and hence to consent to their acts of aggression against you. Arguably, if you believe that what someone is doing to you is completely morally wrong, you can't really *consent* in any permission-implying sense to being his victim. Yet, once this distinction has been made, it is not clear that there is any support for I: if a combatant knows that he's on the just side, he will merely accept his role and distinguish himself from noncombatants but not consent in any permission-implying sense to being the object of unjust attacks.

A second reason why premise I is problematic is that it's simply false that signing up for military service carries with it a commitment to fight in any war whatsoever. If, to take a funny imaginary case that Jeff introduced in his lecture last Friday, you joined the military under President Gandhi and find yourself still in the military when the balance of power shifts and President Stalin takes the helm, it is completely bizarre to claim that the all the risks you accepted in joining the military under Gandhi include the possible risks that you now incur under Stalin.

A third reason why premise I is problematic is that it rests on a blatant overgeneralization. It may be true that *some* potential combatants who joined the military in times of peace commit themselves to fight for their country even in fairly questionable circumstances. But consider the people who join the military only when their country has been clearly unjustly attacked. It is false to say that these people commit to fight in any war whatsoever; otherwise they would have joined the military earlier in life. As Jeff puts it: "These combatants are very much like a man who is suddenly compelled to defend himself against an unjust attacker. Such a man has no reason to waive his right not to be killed and there is no reason to suppose that he does so. Imagine the situation of one of the numerous young Polish men in 1939 who, on learning that the Nazis had invaded their country, rushed to enlist. It seems absurd to suppose that by enlisting they understood themselves to be waiving their right not to be killed, or to be granting the Nazis permission to kill them."[48]

---

[47] Ibid., p.57.
[48] Ibid., p.53.

So, altogether, there seem to be reasons to reject the I-IV argument, and to deny that the 'boxing match' model offers any reason to accept the Moral Equality of Combatants.

4. How Weighty are the Unjust Combatant's Excuses?

I think we're left without any convincing reason to accept the Moral Equality of Combatants in its fullest generality, at least in the case of defensive wars with clearly unjust sides. So far, though, we've been conceding that although some of the factors that defenders of the Traditional Theory bring up – e.g., limited information and duress – may not function as justifications for the acts of unjust combatants in any interestingly deep sense, they *do* function as excuses. I think this is partly as it should be, because everyone should grant that limited information and duress are sometimes good excuses. However, we are left with a serious question about *how weighty* these types of excuse ought to be in the individual case.

Surprisingly, there seems to be a strong case that they may not be very weighty excuses in many cases. One reason for this is that when the risk of acting on limited information or giving into duress is sufficiently weighty, the reason-giving force of the risk trumps the effectiveness of the excuse. To bring this out, consider the following cases, one of which Jeff brought up in class:

> *Dale and the Button.* Dale has decent evidence for believing that, by pressing a certain button, there is a high chance that he will cause someone great pleasure. He realizes that it is consistent with his evidence there is also some tiny chance that pressing the button will end up causing everyone in California inordinate pain for a month and then kill them. After all, he heard someone who is normally just a tiny bit kooky say that there was a small chance that this might happen. Dale presses the button, going with what his evidence seems most to favor, and suspecting that the person he heard was indeed unreliable in this case. Alas, the worst happens.
>
> *Billy's Choice.* A mobster with control over the local police finds Billy walking alone one afternoon and puts him in a bind. The mobster tells Billy that he will have him put in prison unjustly for ten years unless Billy shoots the first person wearing a green scarf he sees and brings the scarf to him. If Billy does this, though, the mobster will use his control over the local police to prevent Billy from being arrested or charged for the killing, and give him some money. Billy decides to shoot the first person wearing a green scarf that he sees, brings the scarf to the mobster, gets his cash and goes home scot free.

I take it that we think Dale and Billy not only made mistakes, but that they are blameworthy for the mistakes they've made. If Dale said, "But, given my evidence, there wasn't much of a reason to think that that would happen", we wouldn't be very inclined to pardon him. If Billy said, "But I was under duress and didn't want to go to prison for ten years", we also wouldn't be very inclined to pardon him, though we might be able to sympathize with his self-interest. But, crucially, both Dale and Billy have what traditionally qualify as sensible excuses: limited information and duress. What these cases seem to show, then, is that when the costs of acting on limited information or giving into duress are sufficiently high, the excusing power of these factors is significantly reduced.

But if that's right, the same reasoning ought to apply to the case of unjust combatants whose only excuses are fear of imprisonment for conscientious refusal to cooperate or having good

but not decisive evidence that they were on the right side. If the impersonal costs of contributing to a war that is in fact unjust considerably outweigh the personal costs of engaging in conscientious refusal or being scrupulous about heeding the fairly small but nontrivial chances of disaster (given one's evidence), these will not function as decisive excuses. Accordingly, unjust combatants may not always or even often be free from blame unless the duress they were under was life-threatening, or the evidence they had that they were on the right side was arbitrarily strong but, as a matter of chance, misleading. If that is right, it poses an even more radical challenge to the foundations of the Traditional Theory than the criticisms we've so far seen.

References

Jackson, Frank. 1991. "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101: 461 – 482.

Kolodny, Niko and MacFarlane, John. Ms. "Ifs and Oughts." http://johnmacfarlane.net/ifs-and-oughts.pdf

McMahan, Jeff. 2009. Killing in War. Oxford: Oxford University Press.

Zimmerman, Michael. 2008. Living with Uncertainty. Cambridge: Cambridge University Press.

1. Towards a Weak Domestic Analogy: Preemptive and Preventive Self-Defense

Today we will be turning to the ethics of preemptive and preventive war. Because I think the best way to think clearly about these issues is *via* a (weak) domestic analogy with individual self-defense against imminent and non-imminent threats, I will first discuss this simpler case.[49] Indeed, I will spend a large amount of time on it, because I see the clearest views about the collective case as close extensions of what we ought to say about the individual case.

I'll start with a bit of quick review of our earlier discussion of individual self-defense. An upshot of that discussion was that, minimally, an innocent person X is permitted to perform a defensive action A against Y when Y is responsible for a threat to X, and A-ing is a necessary and proportionate means for averting the threat. In other words, Y's responsibility for a threat to X was assumed to be a sufficient condition for Y's *being liable* to necessary and proportionate defensive action by X. Notably, in every case we discussed before, the threat was one that had *materialized*: Y was running at X with a knife, Y was pulling the trigger of a gun aimed at X, Y was falling and about to land on X, or Y had pulled a switch that would send a trolley at X.

It seems to me that it's a delicate and difficult task to come up with a good account of what it is for a threat to materialize. But we can at least start with the following working definition, which is probably susceptible to subtle counterexamples that we'll have to ignore:

> *Rough Definition of a Materialized Threat* (MT): An action A done by Y *materializes a threat* to X if and only if ("iff") Y's A-ing initiates a causal chain that, if not interfered with, has a very high probability of leading, *by itself*, to a harm to X.

MT strikes me as good enough to capture the cases we had discussed. The importance of some of the qualifications (like "by itself") will become apparent soon enough.

---

[49] Let me add an important caveat about what I intend by appealing to a "domestic analogy" here, so that no one is confused about the relationship of what I am saying to Jeff's claims in very recent lectures. I think we ought to distinguish two different claims:

> Weak Domestic Analogy (WDA): The moral principles governing conflicts between states are relevantly analogous to the moral principles governing conflicts between individuals.
>
> Strong Domestic Analogy (SDA): States may always be treated in ways that are morally analogous to ways in which individuals ought to be treated.

These claims are partly logically independent: SDA entails WDA, but WDA does *not* entail SDA. SDA is false for reasons that Jeff has been bringing out in his discussion of humanitarian intervention. But the reasons for rejecting SDA stemming from what we ought to say about humanitarian intervention are *not* reasons for rejecting WDA. WDA remains plausible. Indeed, I think Jeff has to believe something *close* to WDA to get some of his arguments for the Revisionist Theory to work. All of this is just to say that when, in the lecture, Jeff claims to be speaking against the domestic analogy, what he really intends to be speaking against is SDA and not necessarily WDA. There might be different reasons for being suspicious of WDA, but we haven't discussed any and, I predict, will not see any in this class.

Now, in thinking about preemptive and preventive self-defense, we face the task of extending the theory of permissible self-defense beyond acts that would avert materialized threats. We will, in particular, be interested in *imminent* and *non-imminent threats*, and the first thing to start with will be definitions of these notions. To get a grip on how these differ from materialized threats, I'll start with a couple of examples:

> *Imminent Threat.* Zane is a freewheeling amateur sharpshooter. He climbs to the top of an abandoned building, and begins to load his gun and attach it to a tripod with the intention of shooting several people down below in a few minutes.
>
> *Non-imminent Threat.* Shane is a disgruntled factory worker. He has planned to put a bomb in the administration office of the factory within a month. He has written about it in his journal. Foolishly, he brought the journal to work, and it was stolen by a co-worker, who read about the plans, and will be telling the administration about them. But Shane just assumes he lost the journal, and keeps his plans alive.

It seems overwhelmingly clear that the agents in these cases are doing highly morally problematic acts. They are not, however, responsible for any *materialized threats* at the time at which the stories have been told. Generalizing from these examples, I'd recommend that we define imminent and non-imminent threats as follows (once again, there are probably counterexamples to these rough definitions, but they should suffice for our purposes):

> *Rough Definition of an Imminent Threat* (IT): An action A done by Y *makes a threat imminent* to X iff Y's A-ing is a *major causally necessary condition* for materializing some threat T to X, and, at the time of Y's A-ing, Y has intentions to perform some further actions in the *very near future* that, together with A, will materialize T.
>
> *Rough Definition of a Non-Imminent Threat* (NT): An action A done by Y *constitutes a non-imminent threat* to X iff Y's A-ing is a *significant causally necessary condition* for materializing some threat T to X, and, at the time of Y's A-ing, Y has intentions to perform some further acts *in the broad future* that, together with A, will materialize T.

IT and NT strike me as correctly classifying the two cases I just considered.

Notice that there are really only two crucial differences between IT and NT. One is the stipulation in IT that Y's A-ing must be a *major* causally necessary condition for materializing a threat. By this, I mean that A-ing is a step which gets very close to materializing a threat, but which may not by itself be sufficient for the threat. Loading a rifle, installing it in a tripod, and aiming it at people without yet touching the trigger is an example: performing this act does almost everything needed to materialize the threat. Clearly, we want the imminent threats that *matter* to be major causally necessary conditions, and not minor ones. Suppose Y plans to shoot ten people shortly after he wakes up one morning. Waking up is a causally necessary condition for doing this: after all, Y can't shoot ten people if he's asleep. But merely waking up doesn't do much to actually advance the materialization of the threat. Hence it seems wrong – indeed, crazy – to say that Y's waking up poses an imminent threat.

The reason why I move from "major" to just "significant" in moving from IT to NT should be fairly obvious. We want to say, in *Non-imminent Threat*, that Shane's planning of the bombing and his making the bomb do something to pose a non-imminent threat. These

acts by themselves do not, however, get Shane very close to materializing the threat. After all, he is going to have to plant the bomb in the factory, and this is something he isn't even going to do for a few weeks. Simply planning and making a bomb do not do nearly as much to advance the materialization of a threat as actually planting it would. This is part, I think, of what makes a non-imminent threat differ from an imminent one, at least in the sense that matters for us.

The other respect in which IT and NT differ is the extremely obvious one – viz., the differing temporal qualifications. By definition, an imminent threat is one that's going to be materialized soon, and by definition, a non-imminent threat isn't going to be materialized soon: hence "very near future" in IT and "broad future" in NT.

The last thing to note is how both IT and NT differ from MT. Recall that part of what I said was crucial for an act to count as a materialization of a threat is that it starts a causal chain that *by itself* has a very high probability of leading to a harm. Clearly, simply loading a gun and aiming it without even touching the trigger do not, by themselves, start such a causal chain. That's why these acts qualify as parts of an imminent threat and not a materialized threat. Something similar goes in an even more obvious way for the acts that constitute a non-imminent threat.

This gives us enough material to start asking questions that need to be answered. One question is *what the fact that Y responsibly poses an imminent or non-imminent threat to people who have not lost their rights not to be harmed does to the rights of Y*. A second question is *what it is to which Y makes himself liable in posing an imminent or non-imminent threat to people who have not lost their rights not to be harmed.*

The second question strikes me as being slightly easier than the first. We can bring this out by imagining a couple of possible types of case. In one type, Y poses an imminent threat of death to some X who hasn't lost his right not to be killed at some time *t*, and *if X does not take defensive action at t, X will not be able to stop the materialization of the threat.* In another type, Y poses a non-imminent threat of death to some X who hasn't lost his right not to be killed at some time *t*, and, just as in the first case, *if X does not take defensive action at t, X will not be able to stop the materialization of the threat.* As it happens, cases of the second kind will be rare: in most cases of non-imminent threats, the presence of a larger interval of time (together with the fact that the causal chain leading from planning to materialization is less advanced) entails that there ought to be more chances to stop the materialization of the threat. But these cases aren't inconceivable. Imagine:

> *The Comatose Islanders.* Gilligan lives on an island with a few other people. He is beginning to hatch his plans to bomb their houses: he is writing out the strategy, and making the bombs. As it happens, the other people will all soon enter comas from which they will only emerge immediately before his bombs go off and injure them severely, though without killing them.

In this case, if the people against whom Gilligan is hatching his plan don't take defensive action before they enter their comas, it will be impossible for them to stop the materialization of his threat. But, before they become comatose, Gilligan's threat is still clearly non-imminent.

With that caveat in mind, what should we say about Y's liability to defensive action by the chaps who haven't lost their rights in cases of these two types? Well, for starters, I find it not at all counterintuitive to suppose that Y is definitely liable to defensive action in both types of case. Exactly what *kind* of defensive action Y makes himself liable to in these types of case seems much less clear. Would X be permitted to kill Y in either type of case, if that really were the *only way* of preventing his threat from eventually materializing, and X knew this fact? I think it's pretty clear that X wouldn't be blameworthy for doing so in such a case. Moreover, I think it's pretty clear that X has a *good reason* to do so. But I am not so clear on whether this reason is a *sufficient reason*, one that would actually make it *permissible*, full-stop, and without qualification, for X to kill Y. Honestly, though, my intuitions are just very vague here, and I wonder what the rest of you think about these types of case. Is X morally permitted to kill Y?

Perhaps what should be said is that there is a completely general issue here having to do with the comparative weight of the proportionality and necessity conditions that figure in any plausible theory of the ethics of self-defense. For the *same problem* arises even for materialized threats. Suppose that Jones is about to hack your arms off and then gash you all over the place, but leave you sufficiently intact to survive, that you are completely innocent, and the only defense you have is to pick up and use a flamethrower which, by chance, was left behind a rock, and which Jones can't see. Since we can stipulate that it would certainly kill him, using the flamethrower is not a proportionate means of self-defense. Yet to say that you have *no reason* to fight back simply because of your lack of more proportionate resources is odd. To claim that you're *permitted* to torch him is not *perfectly plausible*, but surely you have *more reason* to fight than to let him radically disable you and torture you. Moreover, surely the *fact* that you have more reason to do this than to do any alternative (since there's only one) entails that it's awfully close to permissible.

So, it seems to me in general that the degree of strength of the proportionality condition in morally constraining your behavior will depend to some degree on your options: here, I think there's a not completely insane view on which you act *barely* permissibly, and the reason why is that you've been backed into a corner by the limited options. Perhaps, however, what we have here is really a confusion of permissibility and excusability. In any case, the core point is that the explanation of the difficulty of getting a sound moral verdict about cases of the second type we've been discussing may have nothing in particular to do with the fact that that case involves a non-imminent threat that can only be averted by non-proportionate means: instead, it has to do with certain types of dilemma that are generated when your options are so constrained that the *necessary* means for averting the threat are *inevitably less than fully proportional*.

Let's return to the first question, which is what consequences the fact that someone poses an imminent or non-imminent threat has for whether that person has lost any rights. As Jeff pointed out in the lecture, there are laws of *conspiracy* that apply to cases like my original *Non-imminent Threat*. While law and morality can come apart, often the law does reflect people's moral intuitions, and here one might take it to be reflecting an intuition that conspirators have indeed lost some of their rights. However, at least as far as my limited understanding goes, people who conspire to commit some crime are not punishable to the same degree as people who actually commit or are caught in the very act of committing a crime. This suggests that while people who raise a non-imminent threat do have *reduced* rights, they don't

lose all the rights that people who materialize a threat. Similar remarks apply to a lesser degree to people who raise an imminent threat, since there are of course also laws applying to *attempted* crimes.

Note that the two questions I've been considering would be jointly answerable if it were obvious that liability coincides with the loss of negative rights (e.g., rights not to be harmed or killed). This isn't completely obvious, of course, since we've seen that people like Thomson think that the falling fat man is liable to self-defensive action, but he hasn't done anything to lose his rights.

In any case, though, I think there are some crucial upshots of what I've been saying that we can take away that will be useful for our discussion of preventive and preemptive war. One is that defensive action even against imminent and non-imminent threats is sometimes permissible, though exactly what *kind* of defensive action is warranted from case to case remains rather unclear to me. Nevertheless, I think we ought to agree that the amount of defensive action that is appropriately proportional in these cases will be less than that which is justified in cases where people are defending themselves against materialized threats. This may be unsurprising, since proportionality constraints do *not* imply that defensive actions are justified in proportion *just* to the harms being risked, but also in (perhaps less than direct) proportion to the *chance* that these harms will be materialized *given* the elements of risk already in play. And the chance will clearly often be lower with imminent threats in contrast to materialized threats, and lower still with non-imminent threats in contrast to imminent and materialized threats.

2. Risk, Impermissibility and Moral Luck

Before I finally turn to the domestic-analogy-implied consequences of the morality of individual defensive action against imminent and non-imminent threats for the morality of preemptive and preventive war, I want to briefly discuss one further issue that comes up in Judy Thomson's paper "Imposing Risks", which I posted on Sakai. The reason is that I suspect what Thomson has to say will also be very helpful in sorting out what we want to say about the case of war.

To see the relevance of the discussion in that paper, notice that what is perhaps the most morally perplexing aspect of first-person deliberation about taking self-defensive action against imminent and non-imminent threats is that there is a real worry about mistakes. To see this, imagine a somewhat intricate variation on the original *Imminent Threat* case we already considered.

Suppose that there was an impending thunderstorm as Zane loaded his gun and was setting it up on the tripod and practicing his aim without yet pulling the trigger, which he would have waited a couple minutes to do. Amazingly, just as Zane was about to lay his finger on the trigger, a bolt of lighting struck him that didn't harm him, but instead had a remarkable effect on his psychology: it reconfigured his neurons in such a way that all his sociopathic tendencies were simply eliminated, and he suddenly, with a crystal-clear moral intellect, wondered what the hell he was doing with this gun, and decided to disarm himself and go home. Now suppose that, before this minor miracle happened, someone ("Chad") on the ground with a pistol saw Zane taking aim at him and others, and Chad recognized him from

his former career as a policeman as Zane the Criminally Insane, who just got out of a long sentence in jail for some very botched conspiracies to murder. He knew Zane's nature, and knew that the objective chance that Zane would start shooting in moments was approximately 99%. Chad safely ignored the remarkably improbable event that, by sheer luck, ended up befalling Zane. Question: would it have been permissible for Chad to have taken a shot at Zane, given this knowledge of the objective chance?

This is a very hard question. It is, of course, clear that Chad would not have been even the *tiniest bit blameworthy* in taking a shot at Zane. But it isn't so clear that this is what he *ought* to do, or – to actually take the question head on – would have been *permitted* to do. Indeed, there seems to be an argument that this is not what he should do:

A. If Y imminently threatens people with death, but *in fact* will not harm anyone at all and will end up with a cleansed moral character, and there are no other standing offenses for which Y hasn't already been punished, it is not permissible to shoot Y.

B. Given the quirky facts of nature in our case, Zane in fact (though by sheer chance) will not harm anyone and will indeed end up with a cleansed moral character, and Zane has already been sufficiently punished for his past wrongdoing.

C. So, it is not permissible to kill Zane.

There is in fact an inverted correlate of this puzzle (which, roughly following Thomas Nagel and Bernard Williams, I'll call a puzzle of *moral luck*), which is what Thomson discusses most in her paper. Here is a case that brings it out, which I simply quote directly from Thomson, whose neighbor in the case is supposed to be completely innocent:

> It certainly seems plausible to think that the circumstances which now obtain just are not circumstances in which it would be permissible for me to cause my neighbor any harm at all – *a fortiori*, it seems plausible to think that if anyone said to me now, "(1) You ought not to cause your neighbor's death", he would be speaking truly. So far so good. In fact I want some coffee now, and must turn on my stove if I am to have some. If I turn my stove on, I impose a risk of death on my neighbor—it is a gas stove, and my turning it on *may* cause a gas leak into his apartment, or it may cause an explosion, etc. Feeling a surge of moral anxiety, I ask your advice. You say: Absurd. That's a fine stove…and the risk is utterly trivial. So it's *not* the case that you ought not to turn your stove on; that is: (2) It is permissible for you to turn your stove on…. Suppose that, feeling reassured, I turn my stove on. Lo—astonishingly, amazingly—my doing so causes an explosion in my neighbor's apartment, and thereby causes his death. Question: does this show that you spoke falsely when you said (2)? That seems to me to be a very hard question to answer.[50]

As Thomson goes on to point out, there is an argument very much like the (A-B-C) argument – a kind of inverted analogue of it – that applies to this case:

> A*. If you ought not to cause X's death, and A-ing will lead to X's death, you ought not to A.

[50] Thomson (1986: 177 – 178).

B*. In the example, Thomson's turning the stove on (though by sheer and astonishing chance) will lead to the death of her neighbor.

C*. So, Thomson ought not to turn her stove on.

(C*) seems intuitive in retrospect. But if we tell the story the other way, it seems unintuitive.

What are our options in solving this puzzle? Well, there is always the option of going in for an evidence-relative understanding of 'ought'. Still, as emerged from our discussion in the last meeting, I think that this would not ultimately be helpful, because the Kolodny-McFarlane paradox shows that we also need a completely objective, non-evidence-relative 'ought', and it certainly seems like the normative significance of the latter will be greater than that of the former. After all, as we saw before, the evidence-relative 'ought' does not interact with facts about rights, and we *are* interested, in cases of risk imposition, about whether rights are violated. So, this move simply sweeps the problem under the bed, as it were, and doesn't get rid of it.

Perhaps, then, I have been right all along, and that what we ought to say is that (C) and (C*) are both true, and that we simply need to insist on a rather sharp distinction between permissibility/impermissibility and blamelessness/blameworthiness. Thomson and Chad are both utterly blameless for what they do, but they act impermissibly. That's my old line.

I think this line ends up being extremely close to the best one to take, though I spent most of Thanksgiving break thinking about certain unnoticed cases in which it fails egregiously (if anyone is curious, I'd be happy to talk about these cases, which have also convinced me that consequentialism cannot be generally true, though it's very close to true). Even assuming it is right, though, it leaves us with a kind of lingering skeptical angst about preventive and preemptive defense that in fact becomes not entirely trivial when we think about extensions *via* the domestic analogy to the case of war. The fact is that what's improbable is not what's impossible, and there may be cases in which we can never be absolutely confident that, if we engage in preemptive or preventive defense, we will have done the right thing.

Of course, in the individual case, this shouldn't bug us *too* much. Probability is still the guide to life, even if it leads us astray sometimes, and we certainly can't be blameworthy if we do what is most probably right. Surely Chad shouldn't feel like a monster if, after the fact, some nearly omniscient meteorologist-cum-neurologist tells him that Zane was about to get zapped by the lightning of virtue. But things are a little different in the collective case, I think. The reason is that, with war (especially with things like preemptive or preventive nuclear strikes!), a hugely greater number of lives are often at stake. And when risk increases like this, it becomes less reasonable for us to act *even on the assumption that the chance that we're wrong about the rightness of our act is quite low*. There is a general result here which we can roughly state in the following way:

> *The Risk-Probability-Reasonableness Law* (RPR): Let P be the schematic claim that A-ing has an objective chance *n* (measured in [0, 1]) of being a necessary means of averting some exceptionally serious unjust threat T. And suppose that we fix *n* at a pretty high value, but less than 1 – say, .85. As the impartial cost of being wrong about P goes up, the degree to which it is morally praiseworthy to treat P as a reason

> for A-ing goes down. When the cost is extremely high, it may be simply morally blameworthy to treat P as a reason for A-ing unless *n* is arbitrarily close to 1.

I believe that RPR is the central reason why we cannot *directly* generalize what we want to say about individual self-defense against imminent and non-imminent but serious threats to what we want to say about preemptive and preventive war.

Why? Because the costs of being wrong in war are usually greater: just consider the case of whether to use a preemptive nuclear strike. Of course, if we could always know that $n = 1$, the morality of preemptive and preventive war really would structurally collapse into the morality of self-defense against imminent and non-imminent but serious threats. Moreover, the importance of RPR may wash out a little bit, since, in computing the weight of the impartial costs needed to make RPR more mathematically precise (so that we can measure *how much less* acceptable it gets to treat P as a reason for A-ing), we have to include the costs to *ourselves* on the other side: after all, we might be threatened with a nuclear strike. Still, the point is that there is a difference – essentially in degree and not in kind – between the individual and collective cases.

3. Could There Be Just Preemptive or Preventive Wars?

I finally turn a little more explicitly and systematically now to a discussion of the domestic analogy that comes out of what we've said about self-defense against imminent and non-imminent threats. For the sake of simplicity, I will at first try to simply set aside the concerns that flow from RPR by thinking about the easy cases where we can know that $n \approx 1$. When we fix this parameter, there are some reasons for thinking that both preemptive and preventive war could, in some very limited cases, be morally permissible. Since, however, we'll see that this is an extremely unrealistic assumption, the conclusion we'll probably end up with is that the real-world cases in which preemptive or preventive war could be just will be vanishingly few.

What would these cases look like? Take preemptive war first. A preemptive war involves state defense against an imminent threat. This, by a simple generalization of the earlier definition IT, is just a case in which some state Y is engaging in acts that are *major causally necessary conditions* for materializing some big threat T to some state X, and, at the time when Y is performing these acts, Y has intentions to perform some further actions in the *very near future* that, together with A, will materialize T. An example of some major causally necessary conditions would be: actually having built effective nuclear weapons, having set them up for use, and perhaps actually having them quietly aimed at another state under some kind of flimsy ultimatum.

Here it seems that if state X really had enough information, and the threat really was unjust, state X could be justified in engaging in preemptive action if it really was both necessary and proportionate. Of course, the familiar problems about how to weigh necessity and proportionality against each other will come into play. They will probably be practically much harder to implement, since it certainly isn't *as easy to see* what the effective necessary means would be when one is dealing with a big geographical and cultural separation *as it is to see* what these factors would be when one is dealing one-on-one with some unjust attacker. Otherwise, though, if we set aside the concerns raised by RPR, it seems that preemptive

strikes should be uncontroversially morally permissible when the chances are high enough, and there is enough information. Such cases will, however, probably be rare and hard to identify, so this "uncontroversial" philosophical result has at best limitedly clear practical significance.

This limitation is even clearer in the case of preventive war, though there is a kind of exception worth mentioning at the end that Jeff has usefully brought up. A preventive war involves state defense against a non-imminent threat. This, by a simple generalization of the earlier definition NT, is just a case in which some state Y is engaging in some acts that are *significant causally necessary conditions* for materializing some threat T to some state X, and, at the time of Y's performing these acts, Y has intentions to perform some further acts *in the broad future* that, together with A, will materialize T. An example of some major causally necessary conditions would be actually having planned a nuclear strike for some future date, and having built suitable weapons with which to dispatch that strike.

Here it becomes profoundly unclear how there could be cases that aren't almost *purely imaginative* where using preventive attacks to start a war would be the right response to such a threat, even if the threat is known with certainty to be serious. The reason is that it is difficult to see how such attacks could ever be clearly necessary, though they could be proportionate. It is hard to see what genuine practical barriers could make it the case that one cannot *just wait* until the threat turns from a non-imminent one to an imminent one. Once it's imminent, we've then got a case in which a preemptive attack might be more sensible. Sure, if there really were practical impediments of such a nature that a preventive strike was *necessary* to avert the future threat, then perhaps, setting side the worries raised by RPR, we could claim that such a strike would be morally permissible. The issue is just that it's hard to see how such extreme practical impediments could realistically materialize in many cases of interest to us. For that reason, the ethical status of *initiating* a war with a preventive strike will often not be a positive one.

There is, however, a different kind of case to consider that Jeff has discussed. This is the case of a state's engaging in preventive actions *once it has already begun a just war*. As Jeff has noted, it seems less crazy to think that it could sometimes be permissible, once a just war has already begun, for the people on the just side to incapacitate the military resources of the unjust side to a degree that is not proportionate to the offenses for which the unjust side is already responsible, but instead to prevent further likely offenses from occurring. Why is it that this case is different?

Part of the answer seems fairly clear to me, but perhaps I'm mistaken. Engaging in war with an unjust aggressor who clearly wants to continue to engage in unjust aggression provides one with much stronger reasons to believe that future aggression will actually occur. When we are actually in war, we are in a position that is epistemically (i.e., with respect to what we can know) more like individual self-defense: if we actually justly have ground forces on enemy territory, we have a more direct kind of evidence regarding the probability of future attacks than we could otherwise get. We know about the intentions, character, desires and plans of the other side – even if not in exact detail – in a much more vivid and clear way. This provides us with a better basis for assessing the chances, and, since we may end up with much more thoroughly justified belief that the chances of future attack remain high, the cost

of being wrong will tend to get swamped by the quality of this evidence. So, the consequences of RPR are less problematic.

But there still seem to be some limitations even to this point. We could instead aim to defeat the unjust side in a way that would guarantee that there would be no more attacks for a while, and then seek alternative methods *besides* further preventive *attacks* on our end that would prevent more distant threats from them. In short, a necessity condition still plausibly applies even in this kind of case, and engaging in preventive action within a war may not be necessary. It's just that, if it *is* necessary, we would have better evidence from actually engaging one on one with the other side about the fact that it would be necessary. Or so it seems to me.

4. Predictions of the Traditional Theory on Preventive War

What I've just been saying is intended to be theory-independent. It does, however, fit better with the contours of the Revisionist Theory than with those of the Traditional Theory. As I've been suggesting, although *in practice* preventive war is going to be hard to justify on grounds of liability *via* the analogy with individual self-defense, it should not be *in principle* impossible to justify. There are conceivable cases – probably highly imaginary, but conceivable and *ipso facto* possible – in which the analogy would run. What is interesting is that the Traditional Theory entails that preventive war is impermissible *even in principle*.

What's the reason for this? Well, recall that, on the Traditional Theory, as on any plausible theory, just combatants must satisfy a criterion of discrimination, which claims that it is only permissible to attack legitimate targets. What targets are *legitimate*? You should recall that defenders of the Traditional Theory uphold:

> *Traditional Target Legitimacy:* A target is legitimate if and only if it is not innocent.

Here, "innocent" is being used as a technical term derived from the Latin word *nocentes*, which means "the harming ones". An innocent target in this technical sense is a target that doesn't fall among the *nocentes*: so, it's a target that isn't engaging in any harming. Friends of the Traditional Theory thus often identify the innocent with *noncombatants*. Understood in this fashion, Discrimination and Traditional Target Legitimacy jointly entail the following principle:

> *Noncombatant Immunity*: It is not permissible to attack noncombatants.

But now it just follows as a matter of simple logic that the Traditional Theory condemns the initiation of a preventive war. Pretty much as a matter of definition, the people whom the combatants who initiate a preventive war will be attacking will be noncombatants. Hence, they will be innocent, and so not legitimate.

Make of this what you will. While it's of course not crazy to say that initiating preventive wars will almost always be impermissible, it may be a little strange to make this kind of claim on such simple-minded definitional grounds.

What all this may really suggest is that defenders of the Traditional Theory ought to redefine "innocent" to mean something closer to the ordinary sense of the English word. It would be bizarre to claim that people who are engaging in serious conspiratorial planning of an unjust nuclear strike, but who have not yet harmed anyone, are innocent. Of course, if "innocent" is used in the technical sense, it won't be odd. But that just shows that the technical sense is a defective and irrelevant one. If Traditional Theorists traded it in for something more in line with familiar usage, they would avoid the objection.

1. Humanitarian Intervention and the Failure of the Strong Domestic Analogy

In the last meeting, I suggested that we distinguish sharply between the following versions of the "domestic analogy", the first of which I qualify a bit more than I did last time:

> *Weak Domestic Analogy* (WDA): The moral principles governing conflicts between any states that are each sufficiently cohesive in the relevant collective intentions of their constituents are deeply relevantly analogous to the moral principles governing conflicts between individuals.[51]
>
> *Strong Domestic Analogy* (SDA): States can always be viewed as being morally analogous to individual persons.

As I said, these claims are partly logically independent: SDA entails WDA, but WDA does not entail SDA. WDA is, I think, highly plausible, and we need it to get many good arguments off the ground. SDA, however, is not plausible. The main reason for this arises when we think about why humanitarian intervention ought, at least under certain conditions, to be permissible.

To see this, let me give a bit of background about how SDA probably came to be accepted by some people. Part of the reason is due to the idea that states should, in the ideal case, be viewable as collective agents. The notion of collective agency is not by itself odd, though it is a hard task to provide a satisfactory analysis of exactly what it takes for a lot of agents to count as one collective agent. There are clear enough cases of it: when a happy couple decides, after discussing the reasons in play and agreeing about them, to have a child, it's fair to say that this was *their* decision. The decision doesn't merely reduce to the fact that each of them wanted to have a child for good reasons. If two partners independently want this, but don't talk about it or come to agreement about it as a result of genuinely interpersonal deliberation, it's hard to say that *if* they end up having a child *just* because of their individual wants, the fact that they had the child was an outcome of *their decision*. Although it was an outcome of their individual *decisions*, that fairly intuitively isn't the same thing.

Notably, as should be pretty evident in this smaller-scale case, there are strong constraints on what it takes for a set of individuals to count as a collective agent. The decisions at which the group members arrive minimally have to be a product of unbiased responsiveness in

[51] By "sufficiently cohesive in the relevant collective intentions of their constituents", I mean that there is minimal agreement among the people who are governed by the state about whether the state ought to act in some way in starting or continuing a defensive or offensive conflict. Just what makes for the relevant kind of minimal agreement here is, I think, hard to determine, but we'll just have to leave this intuitive for our less technical purposes. One thing to stress, though, is that it isn't a necessary condition for there to be this kind of *minimal* agreement that, say, everyone votes in the same way. Why? Because if people really *antecedently agreed on or at least consented to the mechanisms of the voting system*, they would end up having to consent (with some caveats and exceptions, of course) to the corresponding conditions under which a collection of votes gets turned into a decision, even when that decision is one that some might not have wanted. *This* is the sort of agreement (which crucially involves antecedent consent to certain rules for collective decision-making that may not satisfy *everyone's* preferences) that is needed to cash out the aforequoted phrase in a way that would make WDA plausible. I *definitely* don't have any stronger sense of agreement in mind (e.g., everyone votes the same way).

shared deliberation to the reasons that all the members took to be relevant to the question of whether to act in the way that the decisions are encouraging. Were this constraint ever met for some state, it would not be crazy to claim that we can view it as morally analogous to an individual person: with such a unified collective, there plausibly are rights of non-interference. These rights might be defeasible if the aims of the collective are immoral or the like, but when the decisions simply concern what is in this state's interest (assuming that there are no unjust consequences of promoting this interest), it is plausible that such rights will remain undefeated.

SDA is the product of an overgeneralization of this thought. SDA wouldn't be easily falsifiable if any state were a collective agent in the distinctive sense I've been discussing. So, if states always could be treated as collective agents in this sense, then SDA would be plausible. The problem is just that states *can't* always be treated as collective agents. (Indeed, whether a state could *ever fully* amount to a collective agent in the distinctive sense is deeply unclear, though some of the reasons for quick skepticism are undermined by the qualifications I added about the considerable non-stringency of minimal agreement in the footnote above. Even if it couldn't, the existence of collective agency could be viewed as a matter of degree, and some states might be *much closer* to being collective agents than others.)

Why? Well, states are just organized bodies of people under a single government. The concept is thus sufficiently open-ended that people can in principle be citizens of a state while being denied some crucial privileges in governance that ought to be given to all citizens, or, equivalently but more clearly, while being explicitly discriminated against. Laws can obviously be unjustly discriminatory, and they can easily fail to jibe with moral principles. For this reason, there are easily possible cases (and indeed plenty of actual cases) in which some constituents of a state don't get the share of collective decision-making that they deserve. When this happens, it pretty much follows from the constraint on collective agency that I was just discussing that the state which has those constituents isn't even close to being a collective agent in the fullest sense.

If the state isn't a collective agent in this distinctive sense, there is no reason why it should be obvious that it should have the same rights of non-interference as a cognitively normal individual person. This is precisely why the most plausible morality of humanitarian intervention provides a straightforward counterexample to SDA. We can argue as follows:

1. If (i) some representatives of a state are engaging in genocidal acts against even a non-majority sector of its own population, (ii) there is no way to stop these acts without the interference of some further state, and (iii) the people against whom the genocidal acts are directed would consent to this interference for their sake, such interference is permissible.

2. But if states were always viewable as collective agents, (1) would be false.

3. So, states cannot always be viewable as collective agents.

4. If (3), SDA is false.

Since I take it that (1) is so obvious that anything that conflicts with it is false, (2) is too, (4) is a conceptual truth and (3) is a consequence of (1-2), we must agree that SDA is false.

Now, it's important to recognize some built-in qualifications and further issues raised by this type of argument. One question that I think is tricky is sorting out whether we really need a condition as strong as (iii) in (1) to make the argument work. The following seems plausible: if it really would be better for a population sector of some state for another state to interfere on that sector's behalf, and it wouldn't unjustly harm the other sectors of the former state, such interference should be permissible. But if this is plausible, the consent condition represented by (iii) may not be necessary.

There are, however, delicate issues here: perhaps some oppressed sector of the population thinks, for bad but nevertheless quite autonomous reasons, that they deserve to be unjustly discriminated against. If that's right, then assuming that *sector* can be treated as a collective agent, and that collective agents have the same sorts of rights as individual agents, interference in this case might look like a kind of rights-violation to some people. Whether it *should* look this way is another question, but there's at least conceptual space for disagreement here. If that's right, it is less trivially obvious that interference is permissible. I think it still is *even if* there is a rights-violation, since the good achieved by it is just so enormous that it swamps the normative significance of the violation. But I could see some people (e.g., those of a strongly deontological persuasion) disagreeing with me.

Another, perhaps even harder, question is exactly how strong (i) might be made while leaving (1) still perfectly plausible. There is something a tiny bit odd about the thought that interference should be permissible if the violent discrimination is directed against a very, very small sector of the population. Perhaps, though, interference *is* permissible here, but it's just not *as right as* interference in the more extreme case. But I don't really know: it seems like more work needs to be done on this issue.

2. Terrorism and the (In)significance of Intention to Permissibility

Let's turn to a different issue that we've discussed a bit in recent classes – viz., the fact that intention is relevant mostly to considerations of blame and praise and not to considerations of permissibility. The main reason why we'll be bringing this issue up again is that Jeff has been plausibly suggesting in recent lectures that there is a puzzle about how to show that terrorism differs from military acts in a war that foreseeably harm innocents when we start thinking about the fact that someone's intentions in A-ing are typically irrelevant to the permissibility of his A-ing. In a recent paper I posted on Sakai, Jeff summarizes the puzzle (and furnishes some definitions that will be useful for our discussion) in the following way:

> I suggest…we understand terrorism as the intentional harming (usually killing) of innocent people as a means of intimidating and coercing other people associated with them, usually for political purposes…. [E]veryone agrees that terrorism involves intended harm to innocents and most people have seen that feature as an essential part of the explanation of why terrorism is…morally wrong. But if intention does not magnify the moral objection to killing an innocent person – if, that is, an innocent person's right not to be killed imposes no stronger constraint against intentional killing than it does against foreseen but unintended killing – then

> terrorism should be no more objectionable, other things being equal, than military action in war that foreseeably but unintentionally kills innocent people.[52]

Obviously, a flat-footed way of "resolving" this puzzle would be to simply claim that intention really is deeply relevant to permissibility; if so, there would seem not to be much of a puzzle at all. I think this is the wrong move to make (because it cannot really solve the problem), but I think more needs to be said against the stubborn view than I've said in previous meetings.

First, though, let's rehearse a couple of reasons we've already seen for rejecting the stubborn view. One argument comes from the following now familiar passage from Thomson:

> Suppose a pilot comes to us with a request for advice: "See, we're at war with a villainous country called Bad, and my superiors have ordered me to drop some bombs at Placetown in Bad. Now there's a munitions factory at Placetown, but there's a children's hospital there too. Is it permissible for me to drop the bombs?" And suppose that we made the following reply: "Well, it all depends on what your intentions would be in dropping the bombs. If you would be intending to destroy the munitions factory and thereby win the war, merely foreseeing, though not intending, the deaths of the children, then yes, you may drop the bombs. On the other hand, if you would be intending to destroy the children and thereby terrorize the Bads and thereby win the war, merely foreseeing, though not intending, the destruction of the munitions factory, then no, you may not drop the bombs." What a queer performance this would be![53]

Thomson's clearly onto *something* here, but we have to be a little less rhetorical than her to see what it is. Here, as I understand it, is the reasoning that gives force to this passage:

> *The No Navel Gazing Argument*
>
> I. The things that an agent ought to think about in deciding whether to act in some way just are the things that make that act permissible or impermissible.
>
> II. In deciding whether to drop bombs on Placetown, the pilot should not be thinking about his intentions, but rather on the nature of the external outcome that he is going to bring about.
>
> III. So, given (I) and (II), intentions are not among the things that make acts permissible or impermissible.

While I of course think the conclusion of this argument is close to correct, the argument itself is unconvincing because (I) is false. To see this, we should distinguish two claims implied by (I):

> a. If an agent ought to think about some fact X in deciding whether to A, X is something that determines whether A-ing is permissible or impermissible.

---

[52] McMahan (2009: 16).
[53] Thomson (1991: 293).

b. If X is a factor that determines whether A-ing is permissible or impermissible, then an agent ought to think about X in deciding whether to A.

Neither claim is true. To see how both claims could fail, I'll note first that it is orthodox in ethics to think that there is a big distinction between *right-making factors* and *decision procedures*.

This distinction comes up in an obvious way for anyone who thinks that consequences play *some* role in explaining whether an act is right (or wrong) – and, notably, this includes many nonconsequentialists, since what differentiates between them and consequentialists is whether consequences are the *only* factors relevant to rightness (or wrongness), not whether consequences are *some of* the factors relevant to rightness (or wrongness). It is difficult to determine all the consequences of an act that might be relevant to whether it's permissible: any act has an indefinitely large number of consequences, however small, that extend far into the future. Moreover, and more obviously, it is extremely difficult to determine all the consequences of all the *alternatives* to some act, though everyone should grant that the comparisons between the goodness of these consequences and the goodness of the consequences of the act chosen play *some* role in determining the chosen act's rightness. For this reason, although many people grant that an act's consequences (and particularly their normative relations to the consequences of alternative acts) play some role in explaining its rightness (or wrongness), the same people will not say that it's a great idea to think about all of them (or, clearly, about the consequences of the alternative acts available to us): instead, we should use rules of thumb that manage to produce acts with the best consequences in most cases, even if there are exceptions. Many of these rules of thumb will simply be rules of commonsense morality that most of us follow anyway.

The reason why it's coherent to hold this kind of view is that thinking in detail about *all* the consequences of an act, as well as about the comparisons between its consequences and those of the alternatives, would be so time-consuming for cognitively limited agents like us that we'd actually end up being much less likely to perform enough acts with good consequences! In the words of R. E. Bales, who was one of the first to stress the distinction to which I appeal:

> Calculating and comparing the consequences of alternative acts obviously compounds the complexity [of decision making]. Often we have so little information at our disposal, or are so personally involved, that we cannot calculate with any degree of reliability even if we try. Furthermore, calculating and comparing the relative utilities of alternative acts may be very time-consuming. Indeed, in some cases, circumstances may be such that if we attempt to calculate and compare the utilities of alternative acts, we virtually choose one of the alternatives: the familiar case of the drowning man, and the case of a promise to have done a certain thing by a time that is now in the very near future, are examples of this kind of situation. For if we take the time to attempt to calculate and compare the relative utilities of various helping-the-drowning-man acts and various not-helping-the-drowning-man acts, we virtually choose not to help him.[54]

---

[54] Bales (1971: 258).

So, right-making factors like the total consequences of an act can come apart from decision procedures, which include simple rules of thumb that admit of exceptions, like many of the maxims of commonsense morality. It is coherent to claim that consequences are right-making factors even if we shouldn't think about them (or comparisons between them and the multitudinous consequences of the alternative acts available to us) explicitly in deciding how to act, simply because the difficulty of accurately computing them would end up leading to us performing fewer acts with good consequences than we would perform if we abided by simple rules of commonsense morality, which in many cases do produce good consequences.

But if all this is right – and this is a very orthodox thought in the literature – (I) is false, and Thomson's argument threatens to grossly overgeneralize to the crazy conclusion that consequences are often entirely irrelevant. It's not the *rules of conventional morality* that make acts *that are right because they have good consequences* right, but instead just the *fact that they have good consequences*, and if so, and if we ought to be thinking about the rules and not the consequences simply because of our cognitive and computational limitations, then it just follows that there are things we ought to think about in deciding how to act *that are not primary determinants of the rightness of acts*. The rules are just shortcuts, and often what makes them good rules is the fact that following them leads to the best consequences: the following of the rules *considered in isolation from their own consequences* could easily be irrelevant to rightness and wrongness. If so, (a) is false.

The same sorts of considerations show that (b) is false. If it's just too taxing for us to accurately compute the consequences of acts and the comparisons between the consequences of alternatives, and trying to do this actually leads to fewer good consequences than following simple rules of thumb, then there will plausibly be facts that are relevant to whether an act is right that we should not try to think about it much detail in some cases where the computations would be too hard, and where we should instead think about the simple, defeasible maxims of commonsense morality. That's directly inconsistent with (b).

There are other simpler and even more relevant counterexamples to (I) that make clear how the (I-III) argument really could fail. Surely at least *part* of the value of friendship has something to do with the pleasure it provides us; there's a lot more to friendship than this, but this is part of it. It follows that this is part of what makes friendship fitting to desire. Nevertheless, the best way to think about friendship is probably *not* by engaging in self-interested thoughts about the goods that it would get you; it might be fine to view the pleasure you'd get as a nice perquisite, but to make self-interest an overt reason that is proportionate in one's deliberation to the actual effect it has on the value of friendship seems out of sorts with the nature of at least many kinds of friendship: after all, part of what it *is* to be a friend is to treat your friends interests as being as important as your own (at least within certain bounds). Perhaps I'm just being prudish and old-fashioned here. But surely some will find this plausible. If it were, it would be another counterexample to the idea behind (I), because it would show that part of what explains why it's fitting to pursue an end should not figure into the forefront of our thoughts about that end.

The important upshot of all this is just that it casts a lot of at least *indirect* doubt on the *No Navel Gazing Argument*. There are, as I've been saying, many reasons to think that what we ought to think about in deciding how to act may *come far apart* from what explains why acting in that way is fitting, whether morally or prudentially or otherwise. If that's right, one can't

simply say that because it isn't good to think about the structure of your own intentions in deciding how to act, intentions cannot be relevant to permissibility: for there are *many* things that it isn't good to think about in deciding how to act that actually *do* determine whether that act is fitting! So, while I *almost* agree with Thomson's conclusion, I think the argument she most often offers for it rests on assumptions that are just too problematic to do much help in establishing it.

She does have another argument that I think is more effective. The argument is actually in part familiar from discussions we've had about the gap between blameworthiness and impermissibility, but we might as well take a look at precisely what Thomson says about it:

> Here is Alfred, whose wife is dying, and whose death he wishes to hasten. He buys a certain stuff, thinking it a poison and intending to give it to his wife to hasten her death. Unbeknownst to him, that stuff is the only existing cure for what ails his wife. Is it permissible for Alfred to give it to her? Surely yes.[55]

But what exactly is the argument here? I think it's hard to see this as an argument that intention has *no significance at all, ever,* in determining permissibility. The following is not a good argument:

IV. In Thomson's case, the fact that Alfred has a bad intention in giving the stuff to his wife doesn't make it impermissible to give it to her.

V. If adding a bad intention to the causal ancestry of some act A that is otherwise permissible doesn't make it impermissible, that bad intention has no significance in determining whether the person who does A with it acted permissibly.

VI. So, it follows that bad intentions in general have no significance in determining whether the person who does A with it acted permissibly.

The argument is unsatisfactory in a couple of ways. One way in which it is unsatisfactory is that there is an alternative explanation of what is going on in this case that the argument simply doesn't address. We could say that the good effects of A-ing in this case are just *weightier* in determining how Alfred is objectively morally permitted to act than his bad intention. Since the reasons for the act outweigh the reasons against it, Alfred is permitted to perform the act. But it doesn't follow from *this* that there were *no reasons at all* against his doing the act at that time. This is a straightforward fallacy.

Another way in which the argument is unsatisfactory is that the generalization in (VI) just doesn't follow from (IV) and (V). Even if intentions are really are irrelevant to permissibility in some cases, it doesn't automatically follow that they are irrelevant in all cases. We need a lot more illustrative thought experiments than just the one Thomson offers in the quote to arrive at a conclusion that is this sweeping. These points *by themselves* are not intended to contradict her *conclusion*. They are just meant to show that her conclusion doesn't follow from her premises, and that more needs to be done to get the conclusion.

---

[55] Thomson (1991: 293).

As I've been saying, I agree that intention isn't always a very significant factor in determining permissibility. Nevertheless, I have come to think that some further considerations that are closely related to the two I just mentioned do have some force against the very strong claim that intention is *never in any way* at all significant in determining how someone ought to act. Indeed, I think *this* conclusion is false, and that the fact that it's false has interesting implications.

Let me explain why I think the very strong claim that Thomson and many other people want is probably not true. Here is a simple kind of argument to show that the character of a person's intentions has some effect, however small, on how that person ought to act:

> VII. Other things being equal, any agent ought to be committed to being moral.
>
> VIII. To deliberately and autonomously intend to perform an act that you believe is immoral is to fail to be committed to being moral; equivalently, it is necessary for being committed to being moral that you not deliberately intend to perform an act that you believe is immoral.
>
> IX. If, other things being equal, you ought to A, and B-ing is necessary for A-ing, you ought, other things being equal, to B.
>
> X. So, other things being equal, you ought not to deliberately intend to perform an act that you believe is immoral.
>
> XI. So, other things being equal, if you do perform an act that you deliberately and autonomously intend to do because you believe it is immoral, you do something that's impermissible, since it's impermissible to do what you ought not to do.

Now, crucially, I've stuck in "other things being equal" all over the place in this argument. This is not an *ad hoc* thing to do. For we can, I believe, make calculations of the following sort: we can consider what it would be like for someone to do some component of a larger doing, and consider how moral that would be in isolation from everything else the person is doing. Calculations of this sort allow us to determine the normative contribution that some doing makes to the overall rightness or wrongness of a more complex doing, and I think it would be rash to think that there are no such isolable normative contributions: indeed, I think that whether we ought, all things considered, to do something is at least partly a function of the individual normative contributions of all the elements of the thing done. (There are some recent disputes about this assumption ("atomism about reasons"), and there are some limits to its generalizability, but I think it's pretty intuitive, and it's a good place to start.)

This "other things being equal" qualification matters a lot. Obviously, the fact that someone's A-ing with a certain intention would have all-things-considered very good effects on the world makes a contribution to whether that person ought to A with it. But so does the fact that A-ing with that intention would entail that that person would fail to be committed to being moral. For sure, it is often more important to actually *do acts that have all-things-considered good effects on the world* than to *be committed to doing such acts* – and the two can obviously come apart, since we may have false beliefs and misleading evidence, as Thomson's case involving Alfred illustrates. But the fact that *one thing is more important than*

another obviously doesn't entail that the latter has *no importance* whatsoever. That thought is disastrously fallacious and misguided (though common!).

So, the upshot is that I think the argument I just gave actually entails that there is some real normative pressure exerted by our intentions. It may not be as important as the normative pressure exerted by the consequences of our acts. But it's not a completely empty factor. In this respect, I think we end up with a kind of compromise position between the stubborn view that tries to resolve Jeff's puzzle about the difference between terrorism and acts of war that foreseeably but unintentionally kill innocent civilians by saying that intention makes a huge difference, and the strongly opposing view that intentions have no normative significance. That, at any rate, is that I tend to think these days (i.e., what I've thought within the last two weeks!).

3. Related Points about the Impermissibility/Blameworthiness Divide

The points I've just been making lead me to slightly rethink a couple of other issues I've brought up at various points in the class – viz., the full divisibility (in some cases, at least) between impermissibility and blameworthiness, as well as the general tenability of act consequentialism. The first point is a little easier to see than the second, and I'll start with it.

Assume that the first premise of my argument for the normative contribution of intention to permissibility is true. If it is, I think it just directly follows that an act cannot be both all-things-considered morally blameworthy and yet have no moral reason at all counting against it that could in principle make a contribution to its permissibility. For it *is* true that if you deliberately and autonomously intend to do some act you believe is immoral, you are blameworthy *and* that you do something you have *some* reason not to do *for the very same reason why you're blameworthy*. So, what makes for moral blameworthiness and what makes for certain moral reasons against acts is the same thing. So, an act cannot be both blameworthy and *fully* supported by moral reasons. And since moral reasons *just are* the things that determine whether an act is permissible, it follows that an act cannot be both blameworthy and not *for the same reason* have some fact making a negative contribution to the business of weighing pros and cons that determines permissibility.

Of course, this piece of reasoning entails nothing about *how weighty* this reason is. As I explained in the last section, when it comes to cases in which performing an act with an intention that is blameworthy has good consequences, the moral reasons provided by goodness of the consequences could easily outweigh the reason provided by the norm that stated in VII that one ought to be committed to acting morally. But, in principle, this may not always be true: the reason flowing from that norm of commitment might be strong enough to actually make it impermissible to act with the intention, in which case we would have a case where blameworthiness and impermissibility really are genuinely separable.

4. When Act Consequentialism Might Deeply Fail

If this could happen, it would make room for a very serious and fundamental counterexample to act consequentialism, one that obviously cannot simply be explained way by insisting on the distinction between blameworthiness and impermissibility (since that would beg the question).

Interestingly, I think it *does* happen. To see this, though, we have to ask a question: what is it for something to be valuable? I think the answer to this question has the following form:

> *The Fitting Responses Theory of Value*: For something to be valuable just is for it to provide one with reasons to adopt certain responses, either in one's attitudes or one's actions, toward it. In a slogan: facts about value reduce to facts about what responses in attitudes or acts are fitting.[56]

Now, the kinds of attitudes and actions that are called for by different values may differ significantly. Consequentialists tend to presuppose without any comment that the only fundamental response called for by various values is *production* or *promotion*: they assume that what it is for some kind of thing to be valuable is just for one to have reasons to make it the case that the world has more of that kind of thing (e.g., pleasurable experiences, if you're a hedonistic utilitarian, cases of desire satisfaction, if you're a preferentialist utilitarian, and so on). This assumption is certainly extremely plausible in some cases: what it is for pleasure to be good is mainly for there to be reasons for us to cause there to be more of it, and what it is for pain to be bad is mainly for there to be reasons for us to cause there to be less of it.

But it isn't plausible in all cases. Consider what it is for *persons* (in the biographical sense) to be valuable. Surely the belief that persons are valuable does not commit one to the belief that one ought to cause there to be more persons. The fundamentally correct way to respond to the value of personhood is not to make a lot of babies. You don't say: "Hey, honey, we ought to have kids, since personhood is valuable, and part of what it is for it to be valuable is for us to have reasons to cause there to be more instantiations of it." While that might conceivably have force with some people, it seems a little silly (or at least impertinent) to me. (Perhaps more crucially and relevantly, one doesn't say: "It doesn't matter that I'm about to kill this guy, since my husband and I will be having three kids that we'll be able to sustain and make very happy with the resources we'll take when he's dead. All the value of personhood calls for is a sufficient number of instantiations. And, hell, I'm going to be *increasing* the number of instantiations of happy, autonomous personhood partly by killing him. You ought to *praise* me!")

I'm not claiming here that it's not a good thing for there to be more people (assuming this doesn't also have bad consequences for others, as it might with a world as overpopulated as ours!). Indeed, I think it is good, other things being equal. I just think that the value of personhood doesn't *consist* in the fact that we have reasons of any kind to cause there to be more of it in the world. So, if we want to preserve the thought that what it is for something to be valuable is for there to be reasons to respond to it in certain ways, what response *is* called for by personhood? What fitting response explains why personhood is valuable?

The answer is pretty straightforward: the fitting and relevantly explanatory response is, at least in part, a kind of *respect*. But now there are conceivable arguments for certain deontological constraints that conflict with the very core of consequentialist thinking.

---

[56] Cf. Scanlon (1998) and Ewing (1949) for defenses of this kind of theory. Crucially, this is supposed to be a *reductive* theory: we can reduce facts about value to facts about how we ought to respond in our attitudes and in actions. If this theory is true, the most fundamental concepts needed for normative theorizing in any domain are *deontic* (e.g., ought and reason) and not *evaluative* (e.g., good).

Suppose we grant that making more people is, other things equal, good. There might be cases in which we have a choice between either (i) producing more instantiations of the value of personhood, and (ii) respecting the value of personhood. (In some far-flung but perfectly conceivable case, we might be able to bring a lot of happy people into existence by gruesomely killing a few.) Since respect is what's really called for by this value and not production, this *just means* that there will, other things being equal, be more reason for one to respect the value than to cause there to be more instantiations of it. But that's just to say that it could be permissible to do something other than maximize the good, at least if "maximizing the good" means – as most consequentialists take it to mean – "making it the case that there is as much good in the world as possible". And that's just to say that, at least on the orthodox interpretation, act consequentialism cannot be generally true. There are some agent-centered restrictions on the promotion of value when "promotion of value" is understood in the orthodox sense, and the intuitive appeal of these restrictions cannot be explained away by appeals to the distinction between blameworthiness/praiseworthiness and impermissibility/permissibility. Q. E. D.

References

Bales, R. E. 1971. "Act Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedure?" *American Philosophical Quarterly* 8.3: 257 – 265.

Ewing, A. C. 1949. The Definition of Good. New York: Routledge and Kegan Paul.

McMahan, Jeff. 2009. "Intention, Permissibility, Terrorism and War." *Philosophical Perspectives* 23: 345 - 372

Scanlon, T. M. 1998. What We Owe To Each Other. Cambridge: Harvard University Press.

Thomson, Judith Jarvis. 1986. Rights, Restitution and Risk. Cambridge: Harvard University Press.

Thomson, Judith Jarvis. 1991. "Self-Defense." *Philosophy and Public Affairs* 20: 283 – 310.