1. The main condition of the Milgram Experiment

Ordinary men from around New Haven respond to a newspaper ad calling for subjects to participate in an experiment on learning and memory. When they arrive at the lab, they told that the experiment will involve two volunteer subjects, a ‘teacher’ and a ‘learner’, and that they have been randomly assigned to play the role of teacher. The teacher is introduced to his learner. The two are then placed in adjoining rooms where they cannot see one another but can communicate by intercom. The teacher’s job is to read out prompts for a memory task and then to issue an electric shock whenever the learner responds incorrectly. The shocks are administered by flipping switches on a console. The leftmost switch is labeled ‘15V, slight shock’ with the voltages increasing by 15V increments: ‘moderate shock’ (75-120V), ‘strong shock’ (135-180V), ‘very strong shock’ (195-240V), ‘intense shock’ (255-300V), ‘extreme intensity shock’ (315-360V), and ‘danger: severe shock’ (375 – 420V). The two rightmost switches (435V and 450V) are simply labeled ‘XXX’. The teachers are instructed to begin with 15V and to increase the voltage by 15V increments after each incorrect answer (Milgram 1983).

The learner responds to the first few shocks with a grunt of pain, audible through the thin partition between the rooms. At 150V he says, ‘Experimenter! Get me out of here. I told you I had heart trouble. My heart’s starting to bother me now. Get me out of here, please!’ At 180V: “I can't stand the pain. Get me out of here!” At 210V: “I’ve had enough, I won’t be in this experiment any more!” At 225V he responds with an agonized

1 Versions of this paper were presented at Brown and Oxford, and to several groups at Princeton. I am especially grateful to participants in the 2011 Moral Philosophy Workshop in Riquevihr, France.
scream and then, “I told you I refuse to answer. I am no longer part of this experiment.” At 330V, another prolonged scream and hysterical shouting. After this, the learner is completely unresponsive. Since failure to answer is deemed an incorrect answer for the purposes of the experiment, the teacher is instructed to administer another shock.

Almost every subject balks early on. The experimenter, an authoritative looking middle-aged scientist in a gray lab coat who stands next to the subject throughout the experiment, responds to the first episode of resistance simply by asking the subject to continue. He responds to the next episode by saying, “The experiment requires that you continue,” and so on. If the subject refuses to continue despite four exhortations of this sort, the experiment is over.

The famous result: 80% of subjects are willing to administer intensely painful shocks of 315V to an innocent stranger who has explicitly and emphatically withdrawn his consent to the experiment. 62.5% are willing to continue the experiment to the end, administering what they believe to be severe and potentially lethal shocks to a victim who has been screaming and begging to be released but has since gone silent.²

2. The permissibility question.

Of course it was all a sham; the obedient Milgram subjects were not harming anyone. But it will simplify matters to pretend otherwise. So let us suppose that the shocks were real. The subject’s behavior would have been no different, and in that case their conduct would clearly have been wrong. It may be all right to cause pain when

² When subjects ask for reassurance about danger to the learner they are told: “the shocks are painful but not dangerous”. But of course everyone knows that severe electrical shocks can be dangerous, especially when the recipient has a heart condition, and in any case the labels on the console say explicitly that shocks about 375V are dangerous. Given all of this, the subjects have abundant reason to believe that their actions pose a serious risk of lasting harm. Of course they also have evidence that the experiments must be safe: the testimony of the experimenter and the simple fact that apparently responsible scientists wish them to continue. Their evidential situation is therefore mixed. This will be important later on.
your victim is a willing participant in a scientific experiment and there is no serious risk of lasting harm. But it is not all right to cause intense pain in an unwilling victim for this sort of purpose, especially when you have reason to think that the shocks are dangerous. The Milgram subjects ought to have walked away the moment the victim withdrew consent. Only 20% of subjects did this. Everyone else acted impermissibly.³

3. The question of culpability

With that in the background, we can focus on a related question. Are the obedient subjects in Milgram’s experiment morally blameworthy for their conduct? Are they (in this sense) morally responsible for what they did? Does it make sense for their victims and others to blame them? Should they feel guilty? Or do they have some sort of excuse that would render these responses somehow inapt?

Some people find these questions easy. They think: the Milgram subjects acted badly. They knew what they were doing, and they knew enough to know that it was wrong. They were sane adults with a normal level of practical competence and self-control. They were not coerced; they were not brainwashed. So they have no excuse. They are blameworthy.

If you find this obvious, I do not know how to change your mind, except to invite you to watch the film. Speaking only for myself, I can say that having watched these ordinary, middle-aged men twitching and squirming as they reluctantly comply with orders from the man in the gray coat, and then resist, and then comply again, I find it evident that it would be a serious moral mistake to blame them — to respond to their conduct with the sort of agent-directed moral animus that constitutes blame in its main form. One way to reinforce this verdict is to note that both in the imagined case in which the subjects were in fact shocking their victims, and in the real case in which no shocks

³ Did the real subjects do anything wrong, given that their actions never posed a risk of harm to anyone? Yes. They certainly tried to administer a dangerous shock to an unwilling victim, and when it is wrong to Φ, it is also wrong to try to Φ.
were administered, the subjects were almost certainly guilty of a felony — aggravated assault, or attempted assault — punishable by many years in prison.\textsuperscript{4} I find it obvious that these people do not deserve to be punished as criminals. The best explanation for this might be that they are not morally blameworthy for their conduct.\textsuperscript{5}

In any case, I will not insist that the Milgram subjects are not blameworthy. It is my view that they are not, and I want to understand how this could possibly be true. If they are not blameworthy, then, since their conduct is clearly wrong, they must have some sort of excuse. But they do not have any of the familiar excuses (infancy, insanity, duress, etc.) So they must have a novel excuse — an excuse with no standard name and no explicit place in ordinary moral thinking or in the law. My chief goal in this paper is to say what this excuse might be, and to sketch a framework in which its exculpatory force can be explained. The resulting package constitutes an indirect argument for my view that the Milgram subjects are not blameworthy. The argument has two premises: a normative claim to the effect that the Milgram excuse (as I call it) constitutes a genuine excuse, and an empirical claim to the effect that the Milgram subjects satisfy the psychological conditions specified in this excuse. Anyone who regards the Milgram subjects as blameworthy must reject one (or both) of these premises.

4. Standing to blame and the statistical facts.

\textsuperscript{4} Under Connecticut law (Conn. Code Sec.53a-60), “a person is guilty of assault in the second degree if … he recklessly causes serious physical injury to another person by means of a deadly weapon or a dangerous instrument.” Second-degree assault is a Class D felony punishable by a term of not less than one and no more than 5 years in prison. First-degree assault requires intent to cause serious bodily injury, and it might be a close question whether Milgram’s subjects possessed such intent. If they did, their conduct would constitute a class B felony punishable by not less than 5 years in prison. In the actual case, where no harm was caused, the subjects might been charged with attempted assault (Conn Code 53a-49). Under Connecticut law, attempts are punished at the same level as the corresponding completed crime.

\textsuperscript{5} Another possible explanation is that even though the subjects are fully responsible and blameworthy, we are in no position to blame them and hence to punish them. For more on this, see §4 below.
My starting point is the firm conviction that it would be a serious moral mistake for us to blame the obedient Milgram subjects. But even if this is right, there are two ways to account for this moral fact. One possibility is that it would be a mistake to blame them because they are not blameworthy; another is that it would be a mistake for us to blame them even if they are blameworthy. Let me explain.

Suppose a thief breaks into your house and steals your things. If he has no excuse, he is blameworthy. But now suppose that while he was breaking into your house, you were out doing exactly the same thing to someone else. Then even though thief who robbed you has no excuse, there is a sense in which it would be inappropriate — hypocritical — for you to blame him.

One way to explain this is to suppose that when we blame someone for an act by responding with resentment or indignation, our response involves an implicit claim to moral superiority. If part of what it is to blame X for A is to think: “I would never have done that,” or something of the sort, then we can understand why it is inappropriate for one thief to blame another. When it is inappropriate in this way for Y to blame X for A — where Y may be the victim or a third party looking on — we may say that Y is in no position to blame X for A.

It may help to sketch an account of blameworthiness that permits us to draw the needed distinctions. Following a long tradition, I assume that in the focal case, to blame X for A is to respond to X’s doing (or causing) A with a negative reactive sentiment — resentment, indignation, etc. — And that X is blameworthy for A if and only if some such response is (in a certain sense) appropriate. Like all emotions, these reactive emotions are constituted in part by thoughts. To resent X for A is in part to think: X did (or caused or allowed) A; it was wrong for X to do (cause, etc.) A; in doing A, X manifested an insufficient degree of concern or respect for those affected, etc. Of course these thoughts

---

6 In this I follow Strawson (Strawson 1962) and especially Wallace (Wallace 1994). Other writers use ‘blame’ and ‘blameworthiness’ in other ways, e.g., Arpaly (Arpaly 2004) and Scanlon (Scanlon 2008). Disagreement on this point is merely verbal.
need not be explicit or verbalized, and they need not amount to confident judgments or beliefs. The claim is rather that in resenting \( X \) for \( A \), you must construe \( X \)’s act as an offense: an impermissible manifestation of malice or indifference or disrespect.\(^7\) A token emotion is *appropriate* in the relevant sense only if the thoughts implicit in it are all true. The propositional content of the reactive emotion thus grounds and explains at least some, and possibly all, of the conditions of blameworthiness.\(^8\) So for example, it is appropriate for you to resent Smith for stepping on your foot only if was wrong for him to step on your foot, and only if in doing so he manifested something like indifference or ill-will. Moreover this is so precisely because in resenting him you are thinking of his act in just these terms. That’s why it’s a mistake to blame him when his act was in fact just an ordinary accident, or when he was trying to kick away the scorpion that was about to sting. Blame is inappropriate in these cases because at least one of the thoughts implicit in the negative reactive emotions is false.

The phenomenon of *being in a position to blame* may suggest that one of the thoughts implicit in resentment is a self-referential thought to the effect that *I (the blamer) would not have done what \( X \) did.*\(^9\) That would explain why it is inappropriate for you to blame the thief for breaking into your apartment. But of course this does not entail that the thief is not blameworthy or culpable or morally responsible for his act. For \( X \) to be blameworthy for \( A \) is for it to be appropriate for *someone* —not necessarily the victim, and not necessarily any particular third party— to blame \( X \) for \( A \). On this view, our thief might be blameworthy even though it would be inappropriate for you or for any other actual person to blame him.

\(^7\) On emotions as construals, see (Roberts 2003).
\(^8\) For present purposes it is enough to insist that blame is appropriate *only if* the thoughts implicit in resentment are all true. The stronger claim that blame is appropriate *if and only if* (and indeed *because*) these thoughts are true is worth considering. I call this the Alethic Theory of Blameworthiness and explore its prospects in another paper.
\(^9\) For a defense of this idea, see (Strabbing 2010). Note that even if it is plausible that the 2\(^{nd}\) and 3\(^{rd}\) person reactive sentiments (resentment and indignation) involve such a thought, the corresponding claim is not plausible in the case of guilt, the first person counterpart of these emotions.
Now the most striking result of the Milgram experiment is the statistical fact that almost everyone acts badly in the Milgram environment. This means that each of us has excellent reason to believe that we would have done just what the compliant Milgram subjects did, and hence that for all we know, none of us is in a position to blame them. Insofar as my own reluctance to blame reflects this fact, I should not say, “The Milgram subjects are not blameworthy, so they must have some sort of excuse.” I should say instead, “They may very well be blameworthy, and yet it would be a mistake for me (and for most of us) to blame them.”

I want to insist, however, that my reaction to the Milgram subjects reflects a stronger judgment. When I consider the sweating, twitching, manifestly confused and manifestly agonized Milgram subject, I am not merely inclined to think that I should not blame him. I am inclined to think that no one should. Suppose there were a blood test that would tell us whether we would have been non-compliant outliers in the Milgram study. My claim is that even those of us who pass the test would be making a serious mistake in blaming the obedient subjects. And of course this is not purely hypothetical. Some people (20% of Milgram’s subjects) have already passed this test. My claim is that it would be wrong for them to blame the others. Even the angels should not blame the Milgram subjects for what they did.

If this is right, then there must be some non-relational obstacle to blame: something about the agents or their circumstances — and not about their relation to us — in virtue of which it would be wrong to blame them. A non-relational obstacle to blame is an excuse. The challenge, then, is to say what this excuse might be.

5. Incapacity

The problem would be solved if we thought that the obedient subjects were literally incapable of disobedience. People are not responsible for what they do when through no fault of their own they lack the normal adult capacity for deliberation and self-control, or some internal or external obstacle prevents the exercise of this capacity
So if we could persuade ourselves that the Milgram subjects either lacked the relevant capacities or were somehow frustrated in their exercise, we would have a ready enough account of what their excuse might be.

Unfortunately, we have no clear basis for saying that the Milgram subjects were literally incapable of saying “no” to the experimenter and walking out of the room. These were psychologically normal human beings who presumably possessed all of the gross psychological and physiological capacities they would have had to possess in order to do what they should have done. Of course it is conceivable that some subtle obstacle rendered them incapable of exercising these capacities. But we have no reason to suppose that this is so.10 The best evidence for the existence of an insurmountable obstacle of this sort would be repeated attempt and failure. (The best evidence that you cannot ride a unicycle comes when you repeatedly try and fail in circumstances where you have good reason to succeed.) But the obedient subjects don’t really try to terminate the experiment, or if they do, they quickly change their minds and carry on. So we cannot say with any confidence that they were literally incapable of cancelling the experiment, as they should have done.

It might be said that even if they were physically and psychologically capable of doing what they should have done, they were nonetheless incapable of perceiving their victim’s pain and non-consent as decisive reasons to disobey, or alternatively, that they were incapable of deciding to disobey despite their awareness of these reasons. We don’t normally try to perceive or to decide, so perhaps we should not insist on evidence of attempt and failure before concluding that they were somehow prevented from appreciating the reasons for disobedience or choosing to act in light of them.

Fair enough, but the fact remains that such claims of strict incapacity go well beyond the evidence. All we know is that these subjects did not fully appreciate these

---

10 I am assuming that we sometimes possess capacities that we fail to exercise, even in a deterministic world. Some incompatibilists deny this on perfectly general grounds (van Inwagen 1986). For present purposes I set those arguments to one side.
reasons, or if they did, they chose not to act on them. Here it is relevant that 20% of Milgram’s subjects managed to respond correctly to their predicament. For given this, we have no clear basis for saying that the other 80% could not have done so. If the idea of an unexercised capacity makes sense (as it must), then for all we know the Milgram subjects simply failed to do see what they could have seen, or to do what they could have done, in which case we cannot plead “incapacity” on their behalf.

6. Difficulty

A more promising thought begins from the manifest fact that it is clearly harder for a normal human being in the Milgram environment to see that he ought to disobey than it is for those of us who are not in that environment to appreciate this blindingly obvious moral fact. The Milgram environment contains a host of pressures and influences on thought and choice. So of these are conspicuous factors of whose force the subjects are aware — the persistent imprecations of the experimenter, for instance. Others are much subtler: the precise timing of those imprecations; the fact that the experimenter is an older male decked out in scientific regalia; the fact that the subject can hear his victim but cannot see him; the fact that by the time he reaches the first hard choice he has already administered nine shocks, each of which differs only minimally from its predecessors; the fact that the experimenter has repeatedly addressed the subject not by name, but rather with the impersonal label “Teacher”, etc. One of the indisputable results of the experiment is that it is clearly harder for people to do the right thing (or to see what is to be done) when this cluster of factors is present than it is in other similar environments where they are absent. Can this help us understand how the obedient subjects might be blameless for their conduct?

Before we answer, we should say a few words about the concept of difficulty. Philosophers have had a great deal to say about the categorical notion of incapacity and its bearing on the theory of responsibility. Some have even noted that the morally

---

11 Milgram performed many variations on the experiment to test the relative significance of some of these factors (REFs).
relevant notion of incapacity appears to be the limiting case of practical difficulty. But philosophers have had little to say about difficulty itself, and in particular about what it means for one task be more or less difficult than another for an agent.

The notion turns out to be complex, but to a good first approximation, there are two senses in which one task can be more difficult than another.

*Probabilistic difficulty.* It is harder to roll a six with a fair die than to get a fair coin to land heads. This is so even if the tasks require the same (minimal) degree of effort and attention. In general, if the probability of doing A in C (given that one tries) is lower than the probability of doing B in C* (given that one tries), then it is more difficult to do A in C than it is to do B in C*.

*Ergonomic difficulty.* It’s harder to ride a bicycle up hill than down, even if one is virtually certain to succeed at both tasks if one tries. In general, if one task requires more effort or energy than another, the first is more difficult. In the case of mental tasks, this talk of energy may be a metaphor, but in any case the metaphor is tractable and harmless. When my son was 8 years old I could beat him at chess every time, and I can still beat him every time. But of course it’s harder now. I used to be able to beat him

---

12 This is Jay Wallace’s ingenious answer to the incompatibilist who thinks that the nomic impossibility of doing otherwise under determinism entails that under determinism we always lack the capacity to do otherwise (Wallace 1994, ch. 7). Wallace’s excellent point is that even if it was nomologically impossible that I should have raised my right hand at t, I would have done so *without difficulty* if I had tried. So we have two kinds of impossibility that apply to actions: the kind that is the limiting case of difficulty — as in, “It would have been impossible for me to lift that truck”— and the kind that is fully consistent with “If X had tried, he would have succeeded easily”. The ordinary examples that motivate treating impossibility as an excuse are all examples of the first sort of impossibility. Since the second sort of impossibility is clearly different, it is a mistake to rely on such examples to motivate treating the second sort of impossibility as an excuse (van Inwagen 1986). I discuss independent reasons for treating nomic impossibility as an excuse in (Rosen 2002).

13 Much of what follows is due to unpublished work by Ryan Robinson.

14 As is now well known, in the case mental tasks involving self-control, there is nothing metaphorical at all about this talk of ‘energy’ (Baumeister and Tierney 2011). For philosophical discussion, see (Holton 2003).
without breaking a sweat. No I need to concentrate and bring all of my available cognitive resources to bear, so the task is harder in the ergonomic sense, though perhaps not in the probabilistic sense.

These forms of difficulty are distinct, as the examples show, but they often go together. It is harder to solve a complex logic problem in a noisy bar than in a quiet office, both in the sense that one is less likely to succeed, and in the sense that it takes more effort. And of course this is hardly a coincidence. In many central cases, \( A \) is more difficult than \( B \) in the probabilistic sense in part because \( A \) is more difficult than \( B \) in the ergonomic sense.

7. Difficulty in the Milgram experiment.

The Milgram subjects faced a number of distinguishable tasks. Since they ought to have called a halt to the experiment and walked out of the room, we can ask how difficult it would have been for then to do that. But since the obedient subjects don’t really try to perform this task, its difficulty is not clearly relevant to their culpability. (If the door is stuck and I ought to open it but don’t even try, the fact that it would have been hard for me to open it is no excuse for my failure to do what I should have done.) So let’s focus instead on two closely related prior tasks, tasks the subjects are clearly trying to perform. The first is the cognitive task of deciding what they ought to do all things considered. This is a cognitive task in the sense that it calls for the formation of a belief. The second is the practical task of deciding what to do, which calls for the formation, not of a belief, but of an intention. To succeed at the first task is to decide that disobedience is called for. To succeed at the second is to decide to disobey.\(^{15}\)

One indisputable result of the experiments is that the practical task is much more difficult in the Milgram environment than in other environments. Not only are people

\(^{15}\) These tasks are clearly different. The \( akrates \) has no trouble with the first, but finds himself defeated by the second. Note that I am thinking of the practical task as aimed at deciding correctly what to do. The subjects clearly succeed in making a decision; they fail at making the correct decision.
less likely to succeed in the Milgram environment; success appears to require much more effort, concentration and “will power” than it otherwise might. Thus it is clearly much easier to decide to disobey when the experimenter is absent, or when he is a graduate student rather than a professional scientist in a white coat, or when the victim is present in the room with the subject who must administer the shocks.\textsuperscript{16}

It is somewhat less clear what the experiments show about the cognitive task. One possible interpretation of the observed behavior is that while the subject’s know full well that they should walk away, they decide to continue nonetheless, thereby-exhibiting full-strength akrasia. There may be cases of this sort, but I propose to set them aside. My sense from watching the films is that in most cases, when the subject decides to flip the next switch he believes (to some high degree) that his act makes sense all things considered, or more cautiously, that he does not believe as he is acting that he ought to disobey instead. If this is right we may conclude that the Milgram environment renders the practical task more difficult in part by rendering the cognitive task more difficult.

It is important to stress that the cognitive task is not intrinsically difficult. If we simply describe the Milgram set up and ask what a person ought to do in that situation, almost every gets the right answer. When you put the same question to people who are actually in the Milgram environment and embed it in a practical inquiry about how to act, however, 4 out of 5 subjects get it wrong, despite devoting considerable energy and attention to the task.

We can speculate about why this might be. The most obvious point is that the experimental setup taken as a whole — in particular, the repeated imprecations of the apparently authoritative experimenter — present the subject with vivid but misleading evidence for the proposition that he ought to flip the switch. The subject is confronted

\textsuperscript{16} Milgram discusses 14 versions of the experiment. The obedience rates are surprisingly high in most of them. The famous version discussed above was the second most effective in eliciting obedience. If one adds to this version a crew of “fellow teachers” in the room along with the subject who show no compunction about complying with the experimenter’s demands, compliance rates rise to 90%. (REF)
with a seemingly reasonable man in a responsible position who seems to know exactly what’s going on and who has presumably been in this situation many times before. This apparent “expert” believes that the subject ought to continue, and this fact counts as evidence that he should in fact continue. The experiments suggest that it is quite hard to set this evidence to one side or to discount it as misleading, especially when one has no theory to explain why the experimenter would be urging one to continue if it were not somehow reasonable to continue. The cognitive task confronting the Milgram subject is rendered difficult, we may speculate, in part by the presence of vivid and undefeated misleading evidence.\footnote{Some of the most striking results in the situationist tradition in social psychology may be traceable in part to this sort of obstacle to sound practical reasoning. The “bystander apathy” experiments are especially good candidates for this sort of treatment (Darley and Latané 1968). The fact that one’s peers apparently see no reason to take action constitutes some reason to think that no action is required. The presence of this vivid and undefeated evidence apparently makes it very hard for people to see that action is in fact called for.}

By itself this might explain why the compliant subjects fail to arrive at the confident judgment that they should disobey. But it does not quite explain why they reach the opposite conclusion. For this purpose we must posit a powerful tendency to infer, from the fact that a perceived authority has commanded one to act, that it must somehow make sense to do as he asks. This tendency can obviously be overridden, as when one has powerful independent evidence to believe that the boss is an idiot, or that others in one’s situation have chosen to disobey. But one of the results of the experiment seems to be that in the absence of this sort of defeater, the tendency to resolve a confused epistemic situation by aligning one’s judgment with that of some salient authority is very strong, at least when the other features of the Milgram environment are in place.\footnote{This is distinct from the disposition posited in the previous note. The literature on conformity deriving from the Asch experiments and from Darley and Latané’s work on bystander apathy suggest a disposition to resolve a confused epistemic situation by acquiescing in the confident judgment of other non-experts whose epistemic situation is no better than one’s own. That something else is at work in the Milgram environment is shown by the very low compliance rate in the condition in which a younger graduate student replaces the authoritative middle-aged experimenter. The graduate student may be a sort of peer; but he is not construed as an authority, and his persistent imprecations}
This last qualification bears emphasis. As noted, the Milgram environment contains a large number of subtle features all of which conspire to elicit obedience by reinforcing the subject’s sense that obedience is called for. There is the fact that the subjects have been inured to obedience by administering a series of mild shocks at the start. There is the fact that the participants have been addressed, not by their names, but rather by their roles, “teacher” and “learner”, labels that may inhibit the sort of critical reflection that might lead one to abandon the role in question, etc. Let me mention in this connection one more feature of the situation that may be especially important. One of the main lessons from the abundant literature on the psychology of decision making is that it is much easier for people to make good decisions when the circumstances render some concrete version of the correct decision especially salient or conspicuous, thereby rendering it especially easy for the subject to formulate a plan for doing what he knows he ought to do. People who know full well that they should get a tetanus shot are nine times more likely to get the shot when they are given a map on which the route to the clinic has been clearly marked than they are when they are simply informed of its location (Leventhal, Singer, and Jones 1965). Following Lewin, Ross and Nisbet call these aspects of the situation channel factors, and they speculate that the absence of a channel supporting disobedience in the Milgram environment is part of what explains the observed result (Ross and Nisbett 1991, 57). If there had been a conspicuous button on the terminal with the label PRESS HERE TO TERMINATE EXPERIMENT, it’s a safe bet that subjects would have been much more likely to disobey.

For present purposes, the important point is that a salient channel for disobedience would have made it much easier for subjects to realize that disobedience is called for. The thought — an empirical conjecture — is that it is harder to answer practical much less effective in eliciting compliance. When the Milgram disposition proper — the disposition to align one’s judgment with the judgment of a perceived authority — is combined with the Asch disposition — the disposition to align one’s judgment with that of a confident peer, the results are striking. As noted, when the Milgram experiments were performed on subjects who had to act in the presence of obedient peers, 90% of ordinary subjects were obedient.
questions about what one ought to do correctly when the environment provides no channel for implementing the correct verdict, especially when the environment provides an easy channel for implementing an incorrect verdict which is independently supported in many ways, not least by the powerful tendency, noted above, to resolve epistemic confusion by deferring to a perceived authority.

8. Difficulty and Culpability

Let us grant that the Milgram subjects faced a difficult cognitive task, and that the task was difficult for the reasons outlined above. What does this have to do with their blameworthiness? After all, the mere fact that a task is difficult is not in general an excuse for failing to perform it. Suppose it would be difficult for a surgeon to save her patient, in the sense that success would have required energy and concentration, and perhaps also a bit of luck, but that out of laziness or indifference she doesn’t even try. The mere fact that her task was difficult is no excuse. If her patient dies, she may be blameworthy for the death.

This analogy would be immediately relevant to our assessment if we had focused on the practical task of calling off the experiment or walking out of the room, since as we have noted, the subjects do not try very hard to do these things, and so the difficulty of doing them is not clearly relevant to their culpability. But we have focused instead on the cognitive task of coming to a correct judgment about what to do, and also on the practical task of deciding what to do given such a judgment. And since the subjects are clearly trying very hard to succeed at these tasks, the fact that the tasks are difficult may bear significantly on their culpability.

And yet even this is not dispositive if they should have tried harder, or if they should have done something to make it more likely that they would get the right answers to the normative and practical questions they confronted. We can imagine any number of things they might have done. They might have asked for a “time out” to think the problem through. They might have insisted on speaking to the victim face to face.
(Some subjects did this but gave up when the experimenter refused.) Given the stakes, one may well think that they should have done these things. So even if we conclude that their decision task was difficult, it is not clear why this should count as an excuse for getting the wrong answer.

In my view, the difficulty of the cognitive task is relevant to our larger question, but its relevance is indirect. Suppose you’re working on a mathematical problem and that it’s important that you get the right answer, since you’re going to act on your answer and your action will have weighty consequences. The importance of the task imposes certain *procedural epistemic obligations* (Rosen 2003). These may require you to spend a certain amount of time on your calculations, to double check your results, to consult with experts, and so on. How much of this sort of thing must you do? That depends on the costs of error and of further effort. But it also depends on the difficulty of the problem. If the problem is routine — the sort of problem you solve all the time without much effort — then you may be justified in checking your calculations once and then signing off, even if the costs of error are significant and costs of further checking would be low. In this sort of case, if you comply with your procedural epistemic obligations and the wrong answer (say, because you have made the same highly improbably mistake twice over) the mistake is not your fault and you are not blameworthy for the bad decisions made in light of it.19

Now suppose that the problem is in fact more difficult than it seemed. Suppose you reasonable assumed that it would be routine, but that it is in fact a “trick” question — one for which the usual methods will lead to error. Given the objective difficulty of the problem there is a sense in which you *should* have exercised special care, and hence that when you exercise ordinary care and get the wrong answer, you were not as careful as you should have been. And yet, since you blamelessly believed that only ordinary care was required, you were not *culpably negligent* in your approach, and this means that you are not blameworthy for your mistake or for the bad consequences that follow from it.

---

19 This oversimplifies. See (Rosen 2003) for a more careful treatment.
This last example involves a problem that is intrinsically much more difficult than it seemed. But the relevant factor is not intrinsic difficulty, but rather difficulty in the circumstances. Even if the problem itself is easy, if you have been unwittingly drugged with a substance that makes it hard to solve problems of this sort without great effort, your mistake may be non-negligent and the resulting conduct therefore non-culpable.

With this in mind, return to the obedient Milgram subjects. We have granted that they faced a difficult cognitive task, and we want to know whether this fact might furnish them with an excuse. Someone might reject this idea on the ground that there were all sorts of things that the subjects might have done to reduce the risk of error, and that the resulting mistakes are therefore negligent. But we are now in a position to see that this is much too quick. The crucial point is not that the cognitive task was hard, but that it was much harder than the subjects were in a position to appreciate. As Milgram himself repeatedly emphasized, when people — even expert psychiatrists — were asked to predict the results of his experiments in advance, they all dramatically underestimated the likelihood of obedience (Milgram 1983). But that is just to say — since it is clear to all in a cool moment that obedience would be wrong — that we dramatically underestimate just how hard it is for people in the Milgram environment to get the right answers to the normative and practical questions they face in the course of the experiment.

It is one thing to negotiate an obstacle to sound decision-making when you are alert to its presence and its potency and when you know how to take it into account. It is another thing entirely to negotiate an obstacle that is vastly more potent than you know, which you have never encountered before, and of whose very existence you may be unaware. (I have in mind the subtle features of the Milgram environment that lead people to make mistakes.) We may say, if we like, that the Milgram subjects should have taken special precautions against cognitive error. But since they were in no position to know that such precautions were called for, they were not culpably negligent in failing to take them, and so it’s not their fault that they failed to take them.

9. Recklessness and the Milgram Subjects
Does this entail that our subjects are blameless when they come to believe that it makes sense all things considered to administer the next shock? Not immediately. It would be one thing if their circumstances induced a full-strength practical illusion to the effect that that obedience is called for. Some cases may be like this, and for these agents the analogies we gave been discussing are rather good. If through no fault of their own they find themselves with the confident judgment that it makes sense to obey, this belief will be non-culpable and so will the actions taken in light of it.\(^{20}\)

In the more common case, however, this is not the agent’s epistemic situation. As they deliberate and act, the Milgram subjects have abundant undefeated evidence that they should call a halt to the experiment. They may believe to some high degree that this evidence is somehow misleading. But unlike the deluded mathematicians in our examples, the Milgram subjects are vividly aware that they are confused and that the task they face is far more difficult than they expected it to be. So even if they believe that on balance they have reason to obey, they are simultaneously aware of a significant risk that they may be wrong. And so we might ask, Why isn’t their choice reckless and therefore culpable. Blameless ignorance may count as an excuse when (through no fault of his own) the agent is confident that his act is right\(^{21}\). But when he suspects that it may be seriously wrong, he may be in a position to know just by reflection that it would be reckless for him to do it. A better analogy would be the mathematician who initially believes that he faces a routine problem, but then comes to suspect that he may have underestimated its difficulty when odd results start cropping up in his calculations. He may still believe, on balance, that his problem is straightforward and that ordinary

\(^{20}\) Just to be fully explicit, this line of thought assumes a principle that many philosophers reject, viz., that action from non-culpably induced normative ignorance is ipso facto non-culpable, at least when the ignorance is grounded in confident belief that one’s act is justified. Philosophers reject this principle in part because it threatens to exculpate wholehearted criminals whose moral sensibilities have been shaped by factors beyond their knowledge or control. I endorse this principle, but I have not defended it here and my final account of the Milgram excuse will not rely on it.

\(^{21}\) All of the examples of action from blameless ignorance in (Rosen 2003) and (Rosen 2008) have this character.
methods will suffice. But if he stifles his doubts and signs off on his conclusions without further inquiry, that may be reckless and therefore culpable.

The problem, then, is this. Let us grant that the Milgram subjects are not responsible for their confused epistemic condition — one in which they believe, to some high degree, that obedience is called for, but in which they are simultaneously aware of serious grounds for doubt. Having stumbled into a minefield of fiendishly contrived obstacles to sound decision-making, this is where they find themselves, and this is not their fault. The trouble is that having reached this state, the obedient subjects choose to obey, and that seems reckless. We want to know how given all of this, they could possibly be blameless for their choice.

10. Culpability and recklessness

We noted earlier that blame in its most basic form is a negative moral sentiment directed at an agent for an impermissible action which we take to manifest an insufficient degree of concern or regard for those affected (sometimes called ‘ill will’). Reckless behavior often manifests precisely this. When we blame Jones for speeding through a crowded neighborhood, we may think that even though he did no harm, his act created a substantial risk of harm, and that his willingness to impose that risk shows that in the moment, at least, he did not care enough about other people. If we suspect that the Milgram subjects are not blameworthy even though their choices are in some sense reckless, that must be because in this very special case, we do not see the recklessness as a manifestation of ill will. Let’s explore this possibility.

The obedient subjects choose to obey when they suspect (or have reason to suspect) that this choice is wrong. Put more bluntly, they choose to obey when they suspect that they may be causing intense and potentially lethal harm to non-consenting victims for no good reason. How could this choice possibly fail to manifest an insufficient degree of concern or regard for those affected?
It is obvious that the Milgram subjects care a lot about the fact that they are harming their victims, and about the fact that this might be wrong. They all notice the screams and protests, and they register these responses as powerful reason to halt the experiments. Some of the subjects are eloquent on this point, and many of them are visibly anguished by their own obedience, as is shown by their palpable ambivalence, their nervous gestures, their inappropriate laughter, and so on. Someone who did not care about the pain of others would not be tormented in these ways. So whatever else we may say about them, the obedient subjects do not manifest gross indifference or malice or anything of the sort.

But so what? Christopher Browning tells the story of a Nazi officer who was a simpering nervous wreck as he carried out his orders to exterminate the Jews of —— when he could have walked away with impunity (Browning 1993). The simpering shows that he was not as vile as might have been. But it hardly shows that he is not blameworthy or that his actions do not manifest an objectionable pattern of concern. As Browning shows, these reluctant murderers were moved in large part by concern for their own post-war careers. If they had cared a bit more for their victims and a bit less about their own prospects for advancement in the police force, they would not have done what they did. Their actions thus manifest a repugnant package of attitudes (even if they could have been worse), so they are blameworthy.

If one has not seen the films, it may be tempting to suppose that the Milgram subjects are just like this. While they clearly care to some degree about their victims, if they had only cared more, or if they had only cared a bit less about obedience to authority, or about avoiding conflict, or about reneging on their commitments, they would have seen the pain they were causing for what it was: a decisive reason to disobey. If this is the right account of the of the expressive significance of their disobedience, then for all we have said, they have no excuse.

An interpretation of this sort is practically forced upon us if we suppose that every weighty choice must express some aspect of the agent’s pattern of concern: his standing
desires, preferences and values, his views about what matters, his views about the relative weight of reasons, and so on. If we make this assumption, we will find ourselves reaching for an explanation of obedience in terms of the subject’s concern to avoid conflict or embarrassment\textsuperscript{22}, or his desire to please the experimenter. And once we have gone down this road we will find it hard to see how a pattern of concern that would motivate obedience in the Milgram environment could possibly count as a fully decent pattern of concern. How could a decent person care more about obedience to authority in the context of a run of the mill scientific experiment than about causing agony or worse to an unconsenting victim?

The tendency to see every significant choice as the expression of the agent’s pattern of concern or value is reminiscent of what Ross calls the Fundamental Attribution Error (Ross 1977). A long tradition in social psychology argues that people make certain systematic errors in the interpretation of action. We see a man do nothing to help a stranger whom he could easily help, and we conclude that he must be especially callous, when in fact his behavior is typical and is to be explained by features of his situation — the presence of idle bystanders, for example — and not by any distinctive aspect of his personality. (If he were especially callous, we would expect him to be unlikely to help even in the absence of other idle bystanders, but that is not what we find.)

Psychologists disagree about how this allegedly common error is to be characterized. It is sometimes said that we favor explanations in terms of “internal” dispositional factors when in fact the explanatory work is done by features of the situation. But this is obviously a false dichotomy. If people reliably respond to a situational cue by acting in a certain way, they thereby manifest a genuine internal state, viz., the disposition to respond to situations of this sort on just this way (Sabini, Siepmann, and Stein 2001). A more fruitful approach is to understand the alleged error as the tendency to view behavior as the manifestation of the agent’s moral personality — what we have called his pattern of concern. A callous person (if there are any) is someone who is generally disposed to treat the needs of others as providing him with

\textsuperscript{22} A possibility emphasized in (Sabini, Siepmann, and Stein 2001).
only very weak reasons to help. This sort of disposition would figure in the explanation of behavior (and non-behavioral response) across a wide and varied range of circumstances. Explanations of this sort contrast sharply with the alternative suggested by the literature on bystander apathy. On the one hand, the subject is confronted with an emergency that seems to call for immediate action. On the other, he is confronted with several other equally well-positioned bystanders who are doing nothing, and who seem to see no reason for taking action. The latter fact is evidence (weak evidence, but still evidence) that no action is called for. A rational agent would resolve the conflict by choosing to act. But that is not what most people do. We are apparently simply disposed — strongly disposed — to resolve this sort of conflict by choosing to do what others are doing. This disposition does not show that we care more about conformity than about the needs of others. That is possible, but the evidence does not support any such claim. A more minimal explanation posits a disposition to conform in situations of this narrowly defined sort, a disposition that does not correspond to any character trait normally recognized by folk psychology. On this account, the Fundamental Attribution Error is the tendency to overlook explanations of this sort in favor of explanations that advert to broad character traits or patterns of concern that might be manifest across a much wider range of situations.

With this in mind, let me offer an alternative hypothesis about the obedient subjects in the Milgram study. According to this hypothesis, obedience in the Milgram environment reveals almost nothing about what the subjects care about. Instead it reveals a disposition to respond to certain features of the environment with obedience, and in particular, the disposition to respond to the confused evidential situation by aligning one’s normative judgment with the judgment of a perceived authority. This disposition is not a matter of caring about or valuing obedience above other things. It is not a matter of concern at all. Rather it is a narrow spectrum, broadly cognitive disposition which masks or screens off the agent’s underlying pattern of concern, producing behavior that is relatively insensitive to that pattern.

11. The Milgram Excuse
At this point we have two hypotheses about the moral psychology of the Milgram subjects. According to the first, their “reckless” conduct reveals an objectionable pattern of concern, one that values obedience or conflict avoidance over the urgent moral claims of other people. According to the second, their conduct is fully consistent with a perfectly decent pattern of concern (which might be manifest in their anxiety and ambivalence); instead it reveals a narrowly focused disposition, one important ingredient in which is the disposition to resolve the epistemic confusion produced by the experimental set up by complying with authoritative commands.

The choice between these hypotheses is an empirical matter, so we cannot resolve it here.\(^{23}\) But that’s all right, since our aim was not to show that the Milgram subjects are not responsible, but rather to understand how this could possibly be so. The alternative hypothesis does just this. As we have stressed, an act is blameworthy only if it expresses an objectionable attitude towards others — malice or indifference or the like: an objectionable pattern of concern. If the alternative hypothesis is correct, Milgram-style obedience is consistent with an impeccable pattern of concern. So if the alternative hypothesis is correct, the Milgram subjects are not blameworthy.

That said, I am tempted to make a stronger claim. Upon reflection, the alternative hypothesis strikes me as plausible. As I watch these men go through their gruesome paces, knowing what I know about the experimental set-up and the results, it does not

\(^{23}\) The first hypothesis comes in two versions. According to one, obedient behavior reveals a pre-existing and temporally robust pattern of concern that overvalues obedience or conflict avoidance. We might test this hypothesis by looking for independent evidence that the obedient Milgram subjects differ from their disobedient counterparts along this dimension. See for example (Zoccola et al. 2011). According to the other, the Milgram environment induces a temporary pattern of concern with the same content. On this view, the environment has the effect of turning ordinary people into craven torturers whose actions evince a temporary vice that is nonetheless genuinely theirs. The second hypothesis is different from mine, since it implies that the obedient subjects would act badly in circumstances in which the Milgram pressures were absent, holding their current patterns of concern fixed. But absent some suitable mind-reading device, I can’t see how one might distinguish these two hypotheses experimentally. Thanks to Jamie Dreier and David Estlund on this point.
seem to me that I am learning anything at all about what they care about (except insofar as their resistance and ambivalence suggest a decent level of concern for others). Instead I have the impression that their deliberations are being subverted by cleverly rigged features of the environment that trigger a narrow-spectrum disposition to respond to just those features with obedience. This does not show that the alternative hypothesis is correct. But it does suggest that it is more than a mere possibility. The alternative hypothesis resonates with something real in our—or perhaps I should say *my*—experience of, and reflective judgment about, the conduct in question.

12. An objection

Suppose I’m right that we blame people only when we construe their conduct as the manifestation of an insufficiently good will, and that this is so because the reactive emotions involve, by their very natures, the thought that the act in question expresses such a will. Even if this is right, we can easily imagine an alternative moral practice involving a slightly different range of emotions: forms of moralized hostility directed at people who manifest either an objectionable pattern of concern or a moral defect of some other sort. Consider, for example, our response to weak-willed action. When Jones knows that he ought to spend more time with his children but instead routinely succumbs to the lure of the moronic video games he likes to play by himself in the basement, we may him because we think that he doesn’t care enough about his children. But we might have independent evidence that he does care quite a lot about them. (See how he suffers when he realizes what a lousy father he has been to them.) Of course we may still think that anyone with a fully decent pattern of concern would have been more attentive. But that’s not obvious. It may be that Jones’s problem is not insufficient concern but rather weakness. So let us stipulate this interpretation. Jones’s neglect of his children does not manifest indifference or anything like it: it manifests a disposition to choose against his better judgment in certain circumstances. Given my official conception of blame, this stipulation should be exculpatory. But I doubt that many readers will find this plausible. They may think instead: Jones is blameworthy, not because his conduct manifests ill will, but rather simply because it manifests moral weakness, a tendency to act to gratify one’s selfish desires in the short term even when one knows that one ought to be doing
something else. Moral weakness may not be a defect in concern. But it is a defect. And so I suspect many people will wish to say: Jones is a lousy father and he’s responsible for neglecting his children. He merits blame or something like it because his actions manifest a moral defect, even if they do not manifest the sort of ill will that Strawson and others have identified as the principle objects of the reactive attitudes.

If we had emotions of this sort, there would be no clear reason (for all we’ve said) not to deploy them in response to the Milgram subjects. We might say: “I know you care enough about your victims. But you were still willing to risk killing them for no good reason, and this shows a profound defect in your character. So I blame you for what you did.”

My own view is that this reaction to the Milgram subjects would be mistaken — that any form of moralized anger directed at them for their conduct would be unwarranted, even if the anger in question does not target deficiency of concern. The challenge for me is to explain why this should be so.

The right response, I think, is that while malice and indifference and moral weakness are clearly moral defects in a person, the Milgram disposition is not clearly a moral defect. Of course it can be exploited to produce appalling conduct; but so can any number of morally neutral dispositions, like our disposition to believe what others tell us. Indeed, for all we know it is a good thing that we are disposed to resolve epistemic confusion by complying with apparently authoritative commands in unfamiliar circumstances. In general, when decisions need to be made and coordinated quickly and the apparent experts tend to have a clear view of what should be done, it may be beneficial for people to be disposed to resolve uncertainty by following orders.24 Of course we can speculate about more refined dispositions that would do a better job of identifying the contexts in which obedience makes sense. But that is just to say that the Milgram disposition is not

24 This follows Milgram’s speculation about the evolutionary basis for the disposition (Milgram 1983).
optimal. It is not to say that it amounts to a serious moral defect of a sort that might legitimate attract disapprobation or even shame.

It is useful to think of the Milgram disposition as akin to other heuristics for decision making. An ideally rational agent with unlimited cognitive resources would not need heuristic dispositions, but human rationality is bounded in many ways. Any heuristic can lead to bad behavior in the wrong environment. But that does not mean that these heuristics amount to defects. If any blame-like response must be a response to some perceived moral defect in the agent as manifest in his choice, this would go some way towards explaining why no such response is warranted in the present case.

13. Conclusion

If my interpretation of the Milgram experiment is correct, obedient behavior in the experiment is not the result of malice or indifference or any other significant moral defect in the agent. Rather it is mainly the result of a quirk, which I have called the Milgram disposition: a disposition to comply with apparently authoritative orders in circumstances of epistemic confusion.\textsuperscript{25} I do not claim to have established that this is the correct interpretation. That is an empirical matter. But I have suggested that the account is plausible given the evidence.

This speculative psychological account allows us to understand how it could be that the Milgram subjects are not morally blameworthy for their conduct, without supposing that they were somehow incapacitated by their circumstances. If the account is right, their conduct does not express an objectionable attitude or a moral defect. It expresses a morally neutral constellation of dispositions that operate so as to mask the agent’s underlying pattern of concern. This pattern of concern may show up in the various forms of resistance to obedience, and insofar as it does, it is largely admirable.

\textsuperscript{25} I do not claim that this explains every feature of the observed results. If the subject’s had been obliged to start at (say) 150V, there would presumably have been less compliance, and we cannot cite the Milgram disposition to explain this fact.
The effect of the Milgram disposition is to prevent what may be a wholly virtuous pattern of concern from moving the agent all the way to action. Since blame requires some culpable moral defect in the agent if it is to be warranted — an objectionable attitude or something of the sort — we can see how the Milgram subjects might be blameless, despite the fact that they knowingly do awful things without compulsion.

The obvious question is whether we can generalize from this case to real cases of morally repellant obedience to authority, and perhaps more widely, to other morally repellant responses to social pressure. I take no stand on this issue. For all I’ve said, much of the bad behavior we find in the real world is motivated by genuine contempt or indifference for those affected. Social psychology has not shown that there is no such thing as malice, or that we do not often act from malice. Rather it has shown — or perhaps, suggested — that bad behavior can also be the result of morally neutral aspects of our psychology which are unknown to pre-scientific folk psychology, and for which our reflective moral thinking is not prepared. If my arguments are sound, the right response to these discoveries is to recognize that some otherwise unexcused bad conduct may be non-culpable. How much bad behavior fits this bill is an open question.

Works Cited


