Instances of self-deception are fairly common: A mother confronted with her son’s repeated misbehavior believes that, deep down, he really is a good boy. After writing a novel that meets with poor reviews, a struggling author comforts himself with the thought that his unique talent is simply under-appreciated. Focusing on the sacrifices that he makes for his family, an unfaithful husband assures himself that he is really a devoted spouse and father.

Though the moral problems posed by self-deception have received insufficient attention, the phenomenon seems to pose particularly serious difficulties for Kantian ethics and other autonomous ethical theories. Kant held that acting permissibly consists in acting on an underlying principle or “maxim” that one can will to be a universal law of nature, and he claimed that moral worth can be attributed only to acts that embody such maxims, and which are performed from the motive of duty. Two important problems run as follows: In a broad range of cases in which people deceive themselves about the nature of their actions, their motives, or their underlying principles, Kant’s theory seems to imply that their actions are morally permissible (and

---

1 I am grateful to David Enoch, Bruno Haas, Frances Kamm, Patricia Kitcher, Liam Murphy, Derek Parfit, John Richardson, and William Ruddick for very helpful comments and discussion.

2 In “Self-deception and Autonomy,” Stephen Darwall provides an excellent discussion of the sorts of problems that self-deception can pose for autonomous ethical theories (which he calls “constitutionalist”) and general features of such views that make self-deception especially problematic. See Brian P. McLaughlin and Amelie Rorty, eds. Perspectives on Self-deception. (Berkeley, CA: University of California Press, 1988).
possibly morally worthy), no matter how objectionable they appear. Furthermore, the agents’ self-deception can render them unable to use Kant’s theory to guide their actions properly.

Of course, anyone who acquires false beliefs through certain other means – for instance, by drawing reasonable, but false, conclusions from misleading evidence – runs the risk of performing acts that produce undesirable consequences. One of the purported virtues of Kant’s theory is its claim that such people’s actions are permissible and morally worthy because of their good will, without regard to the unforeseen consequences of their behavior. However, cases in which someone’s self-deception leads her to perform objectionable acts form a special class. Here, claiming that agents’ actions are permissible and morally worthy often seems problematic because the agents should have known better than to behave as they did. That is, the agents’ epistemic irresponsibility accounts for the presence of the false beliefs that contributed to their objectionable behavior.

So the Kantian’s task is to show that we ought to act in ways that minimize self-deception and to deny, or make plausible, the judgment that self-deceivers who commit apparently immoral acts behave in a permissible, or even morally worthy ways. After examining the phenomenon of self-deception in more depth and discussing the ways in which it poses problems for Kant’s ethics, I will explain how the Kantian can meet these challenges.

II

Although the phenomenon is pervasive, self-deception is far from clearly understood. Of course, determining precisely what goes on when someone deceives herself is a difficult task, one that I will not try to accomplish here. However, gaining
familiarity with some of the more promising conceptions self-deception will help us understand both the threat that the phenomenon poses for Kant’s ethics and the range of available Kantian alternatives for addressing the problem. We can make some initial progress by explaining what self-deception is not. I will argue that although self-deception might appear to be an instance of lying, the two phenomena have some important differences.

Suppose an agent truly believes some claim and tries to lead someone else to believe that the negation of this claim is true. Such paradigmatic instances of lying differ from standard cases of self-deception in at least the following respects: First, the liar makes a deliberate effort to deceive someone. However, except in certain extraordinary circumstances, a deliberate effort to deceive oneself seems destined to fail. If an agent is aware of his aim to adopt some false judgment, his awareness seems to rule out the possibility of his taking the judgment to be true. However, if he is not aware of the aim, it is not clear how he could successfully carry out his project of deception. Second, in paradigmatic cases in which a liar tries to deceive someone into accepting a claim, the liar (at least, at some time) truly believes the negation of the claim. However, in the standard cases, someone who comes to accept a false judgment through self-deception need not accept the negation of that judgment at any time.

3 The following may be an example of an extraordinary case: Andy knows that he is likely to forget his activities after a night of heavy drinking. Knowing this, he might drink himself into a stupor, make entries in his diary regarding the night’s events that he knows to be false, and then fall asleep. We can imagine that sometime the next day, Andy reads his diary and comes to believe the false entry. So Andy deliberately deceives himself. However, from the perspective of Kant’s ethics, this case is best understood as an instance of lying rather than an instance of ordinary self-deception. In this sort of case, the agent knowingly acts in ways that will lead him to adopt a false belief at a later time, and, after he engages in the
Third, the success of the lie depends on what Sartre calls the “duality of myself and myself in the eyes of the Other.” That is, a lie cannot succeed unless the deceived agent views the deceiver’s claims and intentions differently than the deceiver himself views them. Since self-deception appears to the product of a single consciousness, it seems difficult to imagine (at least in standard cases) how the self-deceiver could adopt opposing views of both his own aims and the truth of the relevant claims.

Now that we have explored some differences between the lie and the act of self-deception, we can examine a positive sketch of what happens when someone deceives herself, and what sorts of behaviors might facilitate such deception. In Self-deception Unmasked, Alfred Mele suggests that the standard cases of self-deception occur when someone’s desire that some state of affairs obtain (someone’s desire that $p$) causes her to exhibit the following biases in handling the evidence that the state obtains (or fails to obtain): (i) Negative Misinterpretation: The self-deceiver might fail to recognize that some body of evidence counts against $p$, even though, in the absence of her desire that $p$, she would have recognized that the evidence counted against the relevant claim. (ii) Positive Misinterpretation: Someone’s desire that $p$ might lead her to treat certain considerations as counting in favor of $p$, although she would confidently, and correctly, treat these considerations as counting against $p$ in the absence of her desire. (iii) Selective Attending: A self-deceiver might fail to focus on evidence counting against $p$ and focus instead on evidence supporting this claim. (iv) Selective Evidence-Gathering: Someone engaging in self-deception might overlook accessible evidence counting deceptive behaviors and before he adopts the false beliefs, he undergoes some process that distorts his memory. However, the ordinary self-deceiver does not undergo any such process.

against \( p \), but go to great lengths to find evidence supporting this claim. Perhaps people could engage in self-deception by mishandling evidence in other ways, but I will focus on cases in which people arrive at false beliefs by exhibiting these and similar biases as they gather evidence for their judgments.

Perhaps an example will make this account clearer. Suppose that Boyce is a research scientist who spends his career developing a treatment for cancer. Since Boyce has made significant personal sacrifices in order to develop the treatment – neglecting to start a family, failing to spend time with friends, and so on – the thought that his research will not produce any substantial results horrifies him. This anxious desire causes him to mishandle evidence relevant to determining whether his research is successful. For instance, although editors of the most prestigious journals in his field pay little attention to Boyce’s work, he comes to believe that they are not perceptive enough to appreciate the benefits that his research promises to yield (negative misinterpretation). Furthermore, respected researchers at other institutions show little interest in his work. But Boyce comes to believe that his colleagues’ lukewarm response to his research suggests that they are jealous of his promise (positive misinterpretation).

Despite the fact that only a few of Boyce’s patients exhibit any positive response to his therapy, and the vast majority of his patients are responding poorly, he finds it comforting to focus on the former group of patients and distressing to focus on the latter. So he spends a good deal of his free time attending to thoughts of the two or three patients who may be responding favorably to his treatment, and little time attending to thoughts of the patients who are clearly responding unfavorably (selective attending). Finally, Boyce will go to great lengths to discuss his research with the few

---

scientists who think his approach is promising, and he will often avoid reading easily obtainable articles that criticize his methods (selective evidence-gathering).

Some important concerns remain. First, on the sort of account of self-deception I have in mind, the fact that desire causes the self-deceiver to mishandle evidence in various ways does not mean that the self-deceiver deliberately interprets evidence in ways that support his favored conclusion. Rather, beliefs that are acquired through self-deception are produced by some purposive, non-intentional process whose presence presumably has an evolutionary explanation. Mark Johnston describes such a process as a “mental tropism or purpose-serving mental mechanism” that produces a “nonaccidental and nonrational connection between desire and belief.”\(^6\) He suggests, plausibly, that this mechanism serves to reduce the believer’s anxiety that some state of affairs obtains (or fails to obtain).\(^7\)

Although the self-deceiver’s beliefs are generated by a non-intentional mental process, his ability to acquire and sustain such beliefs may be facilitated by behaviors

---


\(^7\) Of course the sort of account that I described is merely one of many types of positive characterizations of self-deception. I will not attempt to survey alternative accounts or try to show that the sort of account I described is preferable to the others. The crucial features of the account I adopted run as follows. (i) The account suggests that self-deception differs in several important respects from instances of lying. This point is agreed upon by most plausible characterizations, and I rely on it to support my claim that an appeal to Kant’s duty not to lie is not sufficient to establish a duty to avoid self-deception. (ii) The account also suggests that acquiring a belief through self-deception is not something that the agent deliberately does – rather it is something that happens to the agent as a result of a non-intentional mental process. A number of the alternative accounts acknowledge this point as well. As I will make clear later on, this presents something like a worst case scenario for the Kantian trying to deal with the problem of self-deception. Any account that somehow makes the acquisition of belief through self-deception more deliberate (while managing to avoid the puzzles associated with viewing the phenomenon as an instance of lying) will only make it less difficult for the Kantian to address the problems I will present.
that he can deliberately adopt or avoid. Such behaviors include the following: (i) The agent might indulge in certain types of fantasy – for instance, Boyce might spend his free hours daydreaming about the respect his colleagues will give him if his research proves successful. (ii) He could also seek the company of others who will encourage his self-deception. For example, Boyce might seek the company of graduate assistants who are confident in his abilities and family members who are eager to see him succeed. (iii) Furthermore, he could avoid the company of those who are likely to discourage his self-deception. For example, Boyce might avoid attending conferences and meetings at which he is likely to encounter respected researchers who are critical of his work. Later on, I will explain how these sorts of activities are relevant to addressing the moral problem of self-deception.

III

Given this characterization of self-deception, we can determine how and why the phenomenon poses a serious difficulty for Kant’s view. The problem is a consequence of a feature of Kant’s ethics that, in some respects, makes the theory both attractive and distinctive. That is, self-deception poses a difficulty for Kant’s view because it is an autonomous ethical theory, according to which the agent’s willing in accord with certain rational constraints constitutes the moral law. As Darwall states in “Self-deception, Autonomy, and Moral Constitution,” autonomous ethical views do not treat certain

---

8 The claim that a self-deceiver deliberately engages in such behaviors does not imply that he engages in them in order to deceive himself.
9 Darwall discusses the relationship between indulging in fantasy and entering self-deception in “Self-deception and Autonomy.”
10 William Ruddick discusses the social dimension of self-deception in “Social Self-Deceptions” from McLaughlin and Rorty, eds. Perspectives on Self-deception.
rules of conduct as fundamental and demand that moral agents try to discover and align their behavior with these rules. Rather, they begin with an independent conception of the character of the ideal moral or rational agent and claim that the moral law is determined solely by the will of agents who conform to the relevant ideal. For such theories, self-deception raises a particularly serious problem. These theories imply that, when someone whose character matches the relevant ideal engages in the relevant forms of self-deception, he might lawfully perform acts that are, at least intuitively, seriously morally wrong. Furthermore, self-deceived agents are able to perform these seemingly immoral acts, while nevertheless exhibiting a character that conforms to the relevant ideal, because they acquired false beliefs in an epistemically irresponsible way. To use Velleman’s words, autonomous ethical theories risk enabling agents who acquire false beliefs through epistemic irresponsibility to “counterfeit morality.”

More precisely, for Kant’s theory, the problem of self-deception can be traced to the following source: Kant claims that an action has moral worth if, and only if, it is both permissible and performed from the right motive, and he claims that acting permissibly consists in acting on principles that conform to certain constraints on rational willing. Kant states in the Groundwork that an action done from duty acquires its moral worth, “not in the purpose that is to be attained by it, but in the maxim according to which the action is determined.” The maxim of an action is the underlying principle that directs someone’s more specific intentions. Onora O’Neill argues that

even routine or thoughtless or indecisive action is on some maxim. However, not all of the principles of action that a particular agent might exemplify count as maxims. For principles of action need only incorporate some description of an agent and . . . of the act and situation, whether these descriptions are vacuous or

---

13 Immanuel Kant Grounding for the Metaphysics of Morals Ak. 399.
brimming with detail. But an agent’s maxim in a given act must incorporate just those descriptions of the agent, the act and the situation upon which the doing of the act depends.\textsuperscript{14}

So a necessary and sufficient condition for an act’s being permissible (and a necessary condition for the act’s having moral worth) is the act’s embodying an underlying principle of the right sort.

Kant identifies the sort of maxim that can serve as the basis of moral worth in his Formula of Universal Law, which reads: “Act only according to that maxim whereby you can at the same time will that it should become a universal law of nature.”\textsuperscript{15} Those actions that embody maxims that cannot be willed into a universal law without contradiction are unworthy, and contrary to duty. By contrast, actions with underlying principles that can be willed into universal law are at least morally permissible and may be morally worthy as well. Whether an act is morally worthy or merely permissible depends on the agent’s motive for acting. We can attribute moral worth to an action if and only if the action is determined by a permissible maxim and the agent performs the action from the motive of duty – in other words, she performs the action because it is right. Kant makes this point in the preface of the \textit{Grounding} when he suggests that “in the case of what is to be morally good, that [an action] conforms to the moral law is not enough; it must also be done for the sake of the moral law;”\textsuperscript{16} and again in the first section when he claims that morally worthy actions must be performed “from duty” and not merely “in accordance with duty.”\textsuperscript{17}


\textsuperscript{15} Kant \textit{Grounding} Ak. 421.

\textsuperscript{16} Kant \textit{Grounding} Ak. 390.

\textsuperscript{17} Kant \textit{Grounding} Ak. 398.
Because of Kant’s conditions for attributing moral permissibility and moral worth, agents who engage in the following forms of self-deception may be able to act in ways that are, according to Kant, morally permissible, and perhaps morally worthy, even though their actions seem seriously morally wrong: (i) First, a self-deceiver can arrive at false beliefs about the motives of her actions. She may, for instance, come to believe that she is motivated to perform an act by her recognition that the act is morally right, when she is really motivated by benevolent feelings, or even by selfish desires. To illustrate, consider a case adapted from one of Kant’s own examples:

**Murderer at the Door**: Becky, hearing a knock at her door, opens it to discover a man wielding an ax. “I’ve come to kill your rich uncle,” the man says. “Is he inside?” Realizing that her uncle is taking a nap, Becky’s desire to inherit some of her uncle’s fortune moves her to respond, truthfully, “Sure, he’s upstairs.” Like most people, Becky wants to be a good person, and she would shudder at the thought that greed could motivate her to sacrifice her uncle’s life. Becky begins to focus on painful truths that she revealed in her past, dutiful actions she performed for other family members, her friends’ claims that she is a good person, and so on. Eventually she comes to believe, through self-deception, that she is moved to reveal her uncle’s whereabouts by her recognition that she has a duty to refrain from telling lies.

(ii) Second, self-deception can lead someone to make an inaccurate assessment of the principles underlying her behavior. This class of cases includes cases in which someone misidentifies the maxim that underlies her actions, as well as cases in which someone incorrectly determines whether her maxim, even if it is correctly identified, passes the categorical imperative test. To illustrate, consider an adaptation of the **Murderer at the Door** example.

**Murderer II**: As before, a murderer appears at Becky’s door and asks whether or not her uncle is inside. Becky reveals her uncle’s whereabouts in order to inherit his vast wealth, and she wonders whether or not her doing so is in accord with duty. Concentrating on previous cases in which she told the truth, and suggesting to herself that she would endure great personal sacrifices in order to avoid telling lies, she comes to believe, through self-deception, that her action embodies the maxim “I will always refrain from telling lies.” However, Becky’s past behavior
and relevant counterfactual truths strongly suggest that the principle that really underlies her action is something like “When doing so poses little risk to me, I will adopt whatever means available to increase my wealth.”

(iii) Third, whether or not an agent correctly identifies her maxims and motives, she may be self-deceived about relevant features of her actions or her circumstances. Her inaccurate understanding either of her own behavior or of the circumstances in which she acts can cause her to behaving in seemingly immoral ways, even though universalizable principles and worthy motives underlie her behavior. For instance, consider the case of the self-deceived doctor that I described above.

**Doctor:** Boyce spent much of his career working on a cure for cancer, and he underwent tremendous personal sacrifices to further his research. Preliminary studies suggest that his proposed new treatment causes somewhat more pain and promises fewer benefits than existing therapies. Moved by his desire that his years of work and sacrifice lead to advances in cancer treatment, he comes to believe, through self-deception, that the therapy he developed will cure his patients’ cancer. Despite the weight of the evidence that his research is a failure, and contrary to many of his colleagues’ advice, he continues to administer the painful and ineffective treatments to patients participating in his clinical studies. He acts on the permissible maxim “Use your talents and expertise to aid people in need.” Furthermore, Boyce administers the treatments because he recognizes that a maxim of non-beneficence cannot be universalized. So he acts from the motive of duty.

In the three types cases that I just described, a self-deceiver commits a seemingly immoral act that Kant’s theory deems permissible, or even morally worthy. Furthermore, unlike those who come to misunderstand their circumstances, or even their own motivations, as a result of reasonable, but fallacious, inferences, the agents in these cases seem to deserve epistemic blame for the beliefs that lead them to act immorally. But there are also cases in which self-deception interferes with someone’s ability to use Kant’s theory to identify and act in accord with her duty. In other words, in the three types of cases that I just described, the problem for Kant’s theory is that it

---

I will assume without argument that an agent cannot universalize this maxim without contradiction.
seems to yield the wrong judgments regarding the permissibility or moral worth of people’s actions. By contrast, in this fourth type of case, whether or not Kant’s theory yields the correct judgment, self-deceivers may be unable, because of their self-deception, to use Kant’s theory to determine how to act rightly.

To illustrate this latter problem, consider another case adapted from one of Kant’s examples.

*Aid*: Carl is a wealthy, miserly businessman trying to figure out whether he should donate any of his fortune to famine relief organizations. The mere thought that he might have a duty to share his wealth with starving strangers causes him considerable distress. So, being a Kantian, he tries to determine whether or not his preferred maxim, “Don’t give aid to people in need” can be universalized. He imagines a world in which no one provides aid, and he comes to believe that he can consistently will that such a world obtain because a man with his wealth and natural talents can flourish without anyone’s help. He overlooks easily accessible evidence that his maxim cannot be universalized and comes to believe through self-deception that his maxim of non-beneficence is in accord with duty. For instance, although he knows that many Kantian philosophers, and even Kant himself, claimed that such a maxim is contrary to duty, he fails to examine their arguments. Furthermore, because he gives the matter relatively little thought, he fails to consider the fact that he cannot guarantee either that his wealth will endure or that he will not develop needs (emotional needs, for example) that his wealth cannot meet.

There is one sense in which this type of case does not pose a special problem for Kant’s view, and a more subtle sense in which Kantians will find this sort of case particularly challenging. One the one hand, the worry that self-deception can interfere with someone’s ability to use a theory to guide their actions is common to all plausible moral theories. For instance, if a similarly situated utilitarian tries to determine whether he should donate to Oxfam, his self-deception might lead him to underestimate the good that his donation would accomplish, or to underrate the precedent effect of his refusal to make the donation. Furthermore, he might exaggerate the contribution to total utility that keeping the money or spending it on luxury items would produce. Or
suppose that our businessman attempts to determine what he ought to do by consulting his moral intuition. In that case, self-deception can lead his intuition astray in the very same ways that it leads him to adopt false beliefs. That is, he might focus his attention on considerations that suggest that his failure to aid people in need is permissible, and he might engage in a less than thorough search for considerations that suggest that his failure to aid is unjustified. Even if the businessman tries to guide his behavior by consulting a list of permissible, obligatory and forbidden acts – something like the Ten Commandments – he leaves room for self-deception to undermine his efforts to act rightly. Someone who is self-deceived can always come to believe that his own case constitutes an exception to some general rule.

So, the fact that self-deception can interfere with someone’s ability to use a given moral theory to figure out what she ought to do seems to be a problem for all plausible moral theories, rather than a problem that specifically plagues Kant’s view. However, there is a further, related worry that is particularly challenging for Kant. Given that self-deception can render someone unable to determine what she ought to do, we can reasonably claim that we have a moral reason to avoid self-deception. That is, we seem to have a moral reason to avoid behaviors that promote self-deception and to engage in behaviors that are likely to prevent us from acquiring or sustaining false beliefs through self-deception. Many of the ethical views I described above possess, at the very least, the resources needed to accommodate a moral reason to avoid self-deception. For instance, the intuitionist can acknowledge that his moral sense or intuition suggests that we have such a duty. Also, suppose that someone’s engaging in self-deception will often result in his unwittingly failing to maximize the net sum of happiness. In that case, the
act-consequentialist can, for this reason, claim that the agent ought to act in ways that are likely to enable him to avoid self-deception.

Furthermore, it is clear that Kant himself was aware of the need for a duty to avoid self-deception. In *The Metaphysical Principles of Virtue* he suggests that the “foremost command of all duties to oneself” is

> know (search, fathom) yourself, not for the sake of physical perfection . . . but for your moral perfection regarding your duty; test your heart – whether it be good or bad, whether the source of your actions be pure or impure . . . Moral self-knowledge, which tries to fathom the scarcely penetrable depths of the heart, is the beginning of all human wisdom. For this wisdom, which consists in accord of one’s will with his ultimate end, requires a man first and foremost to remove the internal hindrances (of a bad will seated within him), and then to try and develop his inalienable original predisposition of a good will. Only descent into the hell of self-knowledge prepares the way for godliness. (Ak. 441)

Unfortunately, it is not clear that Kant’s theory has the resources to accommodate a moral reason to avoid self-deception. It seems that, in order to show that agents who both acquire and sustain beliefs through self-deception are (generally) acting contrary to duty, the Kantian has to identify some maxim on which most (or all) agents who engage in self-deception are acting. Also, he must show that this maxim cannot be willed into a universal law of nature without contradiction. So the task for Kant and his defenders runs as follows: (i) Show that Kant’s ethics has the resources to accommodate a duty to avoid self-deception. (ii) Explain how Kant’s theory can avoid attributing permissibility and moral worth to the actions of self-deceived agents in the kinds of cases I described, or explain why, contrary to appearances, this attribution of permissibility and moral worth is plausible. In the sections that follow, I will explain how Kant can, at least partially, meet these challenges, and I will show how some of these tasks are connected.

IV

14
I will begin with the murderer case in which someone’s self-deception leads her to adopt an inaccurate understanding of the motive that underlies her behavior. I argue that we can avoid attributing moral worth to Becky and her actions through a relatively straightforward application of Kant’s discussion of the opacity of human motives in the *Grounding*. In the second section, he claims that

> there is absolutely no possibility by means of experience to make out with complete certainty a single case in which the maxim of an action that may in other respects conform to duty has rested solely on moral grounds . . . We like to flatter ourselves with the false claim to a more noble motive; but in fact we can never, even by the strictest examination, completely plumb the depths of the secret incentives of our actions. For when moral value is being considered, the concern is not with actions, which are seen, but rather with inner principles, which are not seen. (407)

Three important points from Kant’s discussion deserve emphasis. First, he claims that my underlying motives or “secret incentives” are not merely opaque to those who view me from the outside. Rather, they may be hidden from me as well, and no amount of self-observation or introspection will reveal them to me with certainty. Second, even though my motivations may remain forever hidden from view, there is still a fact of the matter regarding what those motives are. Finally, the attribution of moral worth is tied to what our motives actually are, rather than what I believe them to be. Applying these claims to the murderer case above, we can see how Kant can avoid this aspect of the problem of self-deception. Whether or not Becky believes that she was moved to tell the murderer the truth by her recognition of a duty to refrain from telling lies, her behavior was in fact motivated by her desire to inherit her uncle’s riches. Given the actual motive behind her actions, we are warranted in claiming that her decision to act was not morally worthy.
We can appeal to a similar approach to deal with people who, through self-deception, develop false beliefs about their own underlying principles, like Becky in *murderer II*. Unfortunately, it is not clear whether Kant thought that underlying principles (in addition to motives) could be opaque to the agents; however, I think a good case can be made for the claim that principles may be opaque as well. Suppose, for instance, that I adopt and act on the maxim “When it does not require me to act immorally or to make great personal sacrifice, I will help others achieve their personal projects.” Although I may choose to make an explicit statement of the principle whenever I act on it, or perhaps when I first adopt it, I do not need to formulate the principle explicitly or choose to adopt it (at any time) in order to act on it. Furthermore, it seems that few people formulate their underlying principles explicitly, except in special circumstances. When circumstances demand that we explicitly state our maxims, because, for instance, we wish to perform the categorical imperative test, their precise formulations are not always obvious. Rather they generally have to be discovered through some kind of reflective process, often involving an analysis of our past actions as well as counterfactual claims. The fact that we generally have to engage in this sort of reflection if we wish to discover our maxims helps make room for us to adopt false judgments about them through self-deception.

For instance, suppose that, acting on principle of beneficence described above, I help my father fulfill his dream of opening a pastry shop, and I decide to figure out whether or not my actions are permissible according to Kant’s view. I might initially believe that the principle on which I act is something like “I will help others fulfill their personal projects,” but a little reflection suggests that if my father’s dream had been to rob a bank at gunpoint, I certainly would not have given him any help. A little further
reflection suggests that if my father had asked me to give up all of my own dreams and to devote my life to getting his shop up and running, I would have turned him down. Drawing on this counterfactual evidence, as well as evidence involving my past behavior, my desires and goals, and so on, I can discover that my actual maxim involves helping others fulfill their goals, unless the personal sacrifice is too great or it requires me to violate some duty. However, if I am self-deceived, I may mishandle this evidence and arrive at inaccurate judgments regarding my underlying principles. So, appealing once again to the strategy I used to deal with the previous problem, we can claim that (i) someone may be mistaken about her own underlying principles (ii) there is nevertheless a fact of the matter regarding the identity of those principles and (iii) the permissibility of our actions depends on what our maxims actually are rather than what we believe them to be.

Although this sort of response offers a way out of the first two aspects of the problem of self-deception that I described and seems to remain consistent with Kant’s theory, it may be too strong, ruling out actions that seem permissible. Consider, for example, a case in which someone engages in a thorough and rigorous self-examination and nevertheless arrives at the incorrect judgment that her actions embody a permissible maxim. The kinds of cases that I have in mind are distinct from those in which the agent is simply careless or unresponsive to strong epistemic reasons. In the kinds of cases I am now considering, it simply may be very difficult for someone to determine the identity of her maxim or to figure out whether or not the maxim can be universalized. In such a case, it may seem implausible to judge that her actions are impermissible, given that she put forth considerable effort to ensure that her actions
were permissible, and she simply ended up with an inaccurate understanding of her maxim.

I suggest that even in cases in which someone goes to great lengths to ensure that her maxim is permissible and nevertheless reaches an inaccurate conclusion, her action is impermissible if, in fact, it embodies a maxim that cannot be universalized. After all, although the effort she goes through to ensure that she is acting permissibly is certainly admirable (and might be deserving of moral praise in certain instances), the fact remains that she has adopted and acted on a policy that is not fit to serve as a universal law. Furthermore, denying that such agents’ actions are permissible, in spite of the agents’ efforts to act in accord with duty, merely reflects the fact that morality is complex, that figuring out what we have reason to do and to refrain from doing is frequently a difficult matter.

Unfortunately, even if we accept this sort of response to the first two aspects problem of self-deception, a third problem, which I illustrated with the doctor case, remains. Furthermore, dealing with this third challenge will be considerably more difficult than dealing with the previous two. Although, there do not seem to be any discussions in the *Grounding* that straightforwardly apply to the problem, I will argue for a response that, in spite of its difficulties, remains consistent with Kant’s views. I should emphasize the fact that appealing to the opacity of our motives and principles will not help, because the agent’s understanding of her maxim and motivations is not the source of the current problem. Indeed we can assume, as I did in the *doctor* case, that the agent acts from the motive of duty and that he accurately believes that the principle underlying his behavior can be universalized without contradiction. Rather, the current problem arises from the fact that someone, through self-deception, can
adopt an inaccurate view of his circumstances. As a consequence, his behavior embodies a permissible principle, and it is motivated by the agent’s belief that he acts in accord with duty.¹⁹

An initially promising response to the current problem of self-deception is to base the permissibility and moral worth of agents’ actions, at least in part, on what they would will in appropriately idealized circumstances. This sort of approach appeals to an intuition something like the following: Self-deceived agents like Boyce act on permissible maxims, and they are motivated by their sense of duty, but they would never

¹⁹ Derek Parfit has suggested (in correspondence) that we use the following approach to avoid the judgment that Boyce acted permissibly. I suggested that Boyce’s maxim was “Use your talents to aid people in need.” Of course there is one clear sense in which Boyce acted on this maxim when he administered the therapy. Boyce’s underlying intention was to use his talents to aid people in need, and this is what he took himself to be doing when he administered the ineffective treatment to his patients. However, Parfit suggests that there is another sense in which Boyce fails to act on this maxim, because the facts are not what he takes them to be. That is, Boyce is not really using his talents to aid anyone at all. Since Boyce does not act on a permissible maxim (in one sense), his actions are not permissible. There are a number of problems with this suggestion. First, the account seems inconsistent with the spirit of Kant’s view. Suppose we grant that whether or not an act embodies the maxim that an agent (in Boyce’s case, explicitly) adopts depends on whether or not certain of the agent’s beliefs about his actions and circumstances happen to be accurate. Given that the permissibility of our actions depends on the identity of our maxims, this strategy would make the rightness and wrongness of our actions more dependent on luck than Kant would wish to allow. Second, this view might render the attribution of rightness and wrongness excessively problematic. If Boyce, did not act on the maxim “Use your talents to help others,” in the sense relevant to moral evaluation, we must determine what maxim he did act on in this morally relevant sense. If he did not act on any maxim at all, then his actions are beyond the pale of moral evaluation, which is clearly implausible. However, given the principle that he explicitly endorses, it does not seem any more plausible to attribute some impermissible principle to him – for instance, “Use your talents to subject others to needless suffering.” Third, even if we accept Parfit’s account, we may still be committed to the claim that Boyce acts permissibly if we alter the case appropriately. For instance, suppose that the treatments Boyce administers cause such intense pain that the hospital has to hire additional nurses to comfort his patients. We can suppose that one of these nurses was previously
have committed the acts that they committed if they had acquired accurate beliefs about their circumstances. So we might base our attribution of permissibility, in part, on what the agent would have done if her motives and maxims had remained the same, but her false beliefs about her circumstances had been corrected.

This sort of approach runs roughly as follows: In order to figure out whether an act is permissible, we first identify the maxim that the agent’s behavior actually embodies, and hold this maxim fixed. Then we determine whether, given this underlying principle and keeping other things equal, she would have acted in the same way if her false beliefs about her circumstances had been replaced with accurate ones. In those cases in which the agent’s maxim is permissible and she would have behaved in the same way if her judgments had been corrected, her actions are morally permissible. By contrast, the agent’s actions are not permissible if her behavior embodies a maxim that cannot be universalized or if she would not have performed the actions that she performed in the actual world had her beliefs been corrected. From here, determining whether an action is morally worthy is relatively straightforward. First, we use the procedure I described above to determine whether the action is permissible. We attribute moral worth to the action if, and only if, the action is permissible and it is performed from the motive of duty.

Returning to the doctor example, we can make this approach a little clearer. Boyce’s actions embodied a permissible maxim. Nevertheless, he had false beliefs about his circumstances and the nature of his actions. More precisely, he believed, contrary to unemployed, and that he desperately needed a job to support his family. So, in this revised case, Boyce really does use his talents to help someone in need, and his administering the therapy is permissible.
the best available evidence, that his research promised to yield a cure for cancer and that treatments he administers were actually aiding his patients’ recovery. Suppose that Boyce’s maxim and motives were held constant and that he adopted the true belief that his therapy only causes patients pointless suffering. In that case, given his maxim “Use your talents and expertise to help people in need,” he would not have administered the treatments. As a consequence, we can claim that Boyce acted impermissibly by subjecting his patients to the experimental therapy.

Despite its promise, this sort of approach has at least two faults that make it an inadequate solution to the problem of self-deception. First, I doubt that this sort of strategy can yield, in a non-arbitrary way, plausible judgments regarding which actions are permissible and which are impermissible. For instance, as the following example illustrates, a wide range of cases might obtain in which correcting someone’s misunderstandings about her circumstance would, given her actual principles, result in her acting differently than she does in the actual world. However her actions in these idealized circumstances seem irrelevant to the moral worth and permissibility of her behavior in the actual world. So the idealization approach might implausibly deny permissibility to certain actions.

*Squeamish Assistant:* A hot dog vendor is stricken with a grave medical condition on the side of the road, and a doctor, who happens to be walking by, provides him with emergency care. The doctor asks for assistance from a group of passersby, and Tom volunteers—his actions guided by the permissible maxim, “When doing so involves little personal sacrifice, give aid to others.” Tom usually gets weak in the knees at the mere mention of blood; however, when he decides to assist the doctor, he makes the reasonable, but mistaken judgment that the red stuff smeared across the vendor’s uniform is ketchup. If he had realized that the vendor was covered in blood rather than condiments, he would not have offered

---

20 This is only one example of a general strategy against which I am arguing. I suppose there could be others. The worries I raise are intended to apply to the general strategy.
his help (even given his actual maxim and motives). Rather, he would have been far too nauseous to do much of anything.

If we hope to develop this sort of account into a satisfying response to the problem of self-deception, we face the daunting task of offering some set of criteria that (i) establish when idealized circumstances are relevant to determining the permissibility of someone’s actions (ii) avoid the charge of being ad hoc, and (iii) yield judgments regarding the permissibility of actions that are consistent with our considered intuitions.

Second, even if we set this initial worry aside, another type of problem remains. The idealization involved seems at odds with the Kant’s effort to develop an autonomous ethical theory. Kant argues that the “most direct and infallible way” for me to determine whether my action accords with duty is to ask myself “whether or not I would really be content if my maxim were to hold as a universal law for myself as well as for others.” A central feature of Kant’s theory is the idea that the moral law is determined by the will of actual agents embedded in their actual circumstances. When someone performs an act that is contrary to duty, the agent himself determines that the maxim on which he acted is not fit to serve as a moral law, because the agent himself is unable to will that the principle underlying his action become a universal law. By contrast, according to the idealization approach, the moral law is constituted, not by the will of the actual agent, but by the will of a counterpart with idealized knowledge or capacities.

V

If Kant is able to meet this third challenge posed by self-deception, his response has to be of a very different sort than the responses I have so far described. I argue that the best alternative open to the Kantian is to grant that insofar as they embody

\[21\text{ Kant } Grounding \text{ Ak.403 (emphasis added)}\]
permissible motives and are performed from the motive of duty, the actions of people who engage in the sorts of self-deception I described are permissible and morally worthy. However, we can also identify some other aspect of such agents’ behavior that (at least in most cases) accompanies the permissible action and is itself impermissible. In other words, although people who are self-deceived in the ways I described perform permissible (or even morally worthy) actions, many (or perhaps all) such agents make possible the acquisition or maintenance of the false beliefs on which they act by behaving in morally impermissible ways. This approach brings to light the connection between two distinct problems that the possibility of self-deception poses for Kant’s ethics.

To begin with, as I suggested earlier, the fact that self-deception can interfere with someone’s ability to use Kant’s theory as a guide for action seems to ground a moral duty to try to avoid acquiring or sustaining false beliefs through self-deception. Clearly the Kantian can accommodate such a duty if she identifies some behavior that (generally) makes self-deception possible and embodies an impermissible maxim. Furthermore, as I argued above, Kant’s theory seems to attribute moral permissibility and worth to the actions of self-deceived people who behave in what appear to be seriously immoral ways. So, identifying behaviors on which acquiring and sustaining beliefs through self-deception generally depends and showing that such behavior embodies impermissible maxims will allow the Kantian to identify an impermissible aspect of such agents’ behavior.

One proposal that seems to follow quite naturally from Kant’s theory, but ought to be rejected, runs as follows: We might claim that the very same considerations that ground a duty to refrain from lying also ground a duty to refrain from self-deceiving.
Unfortunately, this sort of approach cannot work since, as I mentioned earlier, both the success of the lie and the success of the Kantian argument that explains why lying is wrong depend on features of lying that self-deception does not possess.

As I argued earlier, the lair cannot succeed unless she exploits a distinction between the way she views herself and her actions and the way she and her actions are viewed by the deceived. Borrowing an example from Kant, suppose that I wish to make a false promise to someone in order to extricate myself from some financial difficulty. Since I have no intention of keeping my promise, the success of my lie depends on the other person’s adopting a different understanding of my intentions than I have. I can only gain the other party’s trust, on which the success of my lie depends, if he views my promise as an honest expression of my intentions. Kant’s explanation of the impermissibility of making a false promise also seems to presuppose what Sartre called the “duality between myself and myself in the eyes of the Other.” 22 Kant claims that the maxim, “I will make false promises in order to extricate myself from difficulty” cannot be willed into universal law without contradiction. If such a maxim became a universal law,

> there would really be no promises at all, since in vain would my willing future action be professed to other people who would not believe what I professed . . . Therefore my maxim would destroy itself as soon as it was made a universal law.” 23

Willing such a maxim into a law of nature would destroy the trust required for such a maxim to succeed precisely because such a law would publicize my intentions to break my promises. In other words, it would eliminate the distinction between the way I (the liar) view my promises and the way they are viewed by those I try to deceive. As Sartre’s

22 Sartre pg. 150
discussion of self-deception suggests, this duality between the deceiver and the deceiver in the eyes of the deceived does not exist in self-deception. Rather, self-deception involves someone’s mishandling her evidence for a given claim, and this mishandling of evidence can occur entirely within a single consciousness.

If an appeal to the self-deceiver’s maxims can ground the claim that people who engage in self-deception generally act impermissibly, the relevant maxim must be of a very different sort than the maxims on which liars act. As I suggested above, the liar makes a deliberate effort to mislead someone. So the principle that his lie embodies will have the following form: “When such and such circumstances obtain, I will deceive someone in order to achieve such and such purpose.” However, in standard cases of self-deception, the agent does not make any deliberate effort to deceive herself. Rather, some anxious desire causes her to refrain from undergoing the sort of thorough reflection needed to ensure that her beliefs are true. So any maxim relevant to an agent’s engaging or failing to engage in self-deception will be concerned, not with deliberate deception, but with the agent’s failure to adopt certain measures to ensure that her beliefs are true.

Since Kant’s duty to refrain from telling lies fails to cover standard cases of self-deception, the Kantian must to appeal to some other duty to show that seemingly immoral behavior resulting from self-deception actually has an immoral component. The simplest and most plausible candidate is a duty to aim diligently at truth in one’s judgments, and the Kantian can ground such a duty in the fact that the following maxim cannot be consistently willed into a universal law: “When I am inclined to accept some claim, I will not take due care to determine whether I have sufficient reason accept the

23 Kant *Grounding* Ak. 403
claim, reject it, or withhold judgment altogether.” (I will call this principle the *maxim of acquiescence*.)

Suppose the Kantian can show that most self-deceivers act on something like the maxim of acquiescence, and that actions that embody this maxim are contrary to duty. In that case, the Kantian can, at least partially, meet both of the challenges I described above. That is, she can show that, on Kant’s theory, we have a duty to act in ways that prevent the acquisition or maintenance of self-deceptive beliefs. Furthermore, she can show that Kant’s theory has the resources to condemn (some aspect of) the behavior of agents who act in intuitively immoral ways, in part, because they acquire false beliefs through self-deception.

Of course, some may worry that we simply cannot establish that engaging in self-deception is contrary to duty. After all, as I pointed out above, ordinary self-deception does not involve any deliberate effort to hide from oneself some truth that one possesses. So it may be difficult to understand how any underlying principle could help

---

24 I should make two clarifications here. First, what counts as “rigorously aiming at truth” may depend on contextual factors – the sort of claim whose truth the agent is trying to determine, the agent’s particular circumstances, and so on. For instance, if I am reminiscing about old times with my college friends, and I want to know whether the Owls won the championship in 1998 or 1999, then determining that one or two friends agree that the Owls won in ’98 may be sufficient for “rigorously aiming at truth.” However, if I am a lawyer trying to determine whether Jones was at the crime scene at the time of the murder, rigorously aiming at truth may involve much more effort. For instance, it might involve my interviewing a number of witnesses, administering lie-detector tests, dusting the scene for fingerprints, and so on. Second, we may worry that a duty to aim rigorously at truth is implausibly strong – that it rules out an agent’s ‘positive thinking’ about her chances of success in uncertain projects. I think the Kantian can accommodate such positive thinkers. As Johnston points out, acquiring certain beliefs increases the likelihood that the believed claim is true. (“Self-Deception and the Nature of Mind” pg. 68) For many positive thinkers, the belief that they will succeed in some uncertain pursuit actually provides (some) evidence for the claim that they will succeed. I think my discussion below will make clear why extreme positive thinkers, who do not work to prevent themselves from acquiring an estimation of their abilities that far exceeds any abilities they could actually attain, are acting contrary to duty.
determine whether or not someone engages in self-deception. Since acting in a way that runs contrary to duty consists in acting on a principle that fails to conform to certain rational constraints, actions that do not embody any syntactically structured principles are not even candidates for being in accord with, or contrary to duty. Suppose we grant, as Johnston suggests, that the actual adoption of a false belief through self-deception results from a non-intentional mental tropism. Even so, actions that I perform both prior to, and after my acquisition of a false belief may determine whether or not I adopt and sustain that belief through self-deception. If these sorts of actions generally accompany instances of self-deception and they embody underlying principles, then someone’s engaging, or failing to engage in self-deception may be subject to moral evaluation on Kant’s view.

To help explain the relevance of the maxim of acquiescence that I described above, we can appeal to Sartre’s suggestion that engaging in self-deception is in some respects analogous to putting oneself to sleep. Of course, falling asleep is not something one can do through a simple act of will. In fact, we often find that the harder we try to fall asleep, the more difficult falling asleep becomes. Nevertheless, a number of behaviors that someone might deliberately engage in or intentionally avoid help determine whether and when she will fall asleep on a particular occasion. On the one hand, she might adopt an attitude of acquiescence, and simply allow sleep to overtake her. Furthermore, she might create an environment in which she is more likely to fall asleep, by playing soothing music, focusing on pleasant memories, reading a boring

---

25 David Enoch raised this worry.
textbook, and so on. On the other hand, she might work diligently to resist falling asleep by avoiding those behaviors that increase the likelihood of her falling asleep or by acting in ways that make falling asleep more difficult. For example, she might concentrate on particularly disturbing ideas, splash water on her face, play loud music, and so on. So, even though we cannot fall asleep by a mere act of will, we can generally fall asleep by adopting an attitude of acquiescence and by creating an environment in which sleep is likely to overtake us. Furthermore, we can often ward off sleep, at least for a while, by working hard to stay awake.

As in the case of falling asleep, the adoption of either of these attitudes, acquiescence or diligent resistance, is relevant to our engaging in, or avoiding, self-deception. Generally, we cannot succeed in hiding some unpleasant truth from ourselves through a simple act of will – that is, though a deliberate effort to deceive ourselves. Nevertheless, many behaviors that embody underlying principles help determine whether or not we acquire false beliefs through self-deception, and whether or not we maintain the false beliefs we have acquired. For instance, if someone adopts an attitude of acquiescence, she will simply accept the judgments that she is motivated by her desires to adopt, without exerting effort to ensure that her beliefs are true. Furthermore, she is likely to engage in behaviors that make it easier for her to sustain the false beliefs. For instance, she could avoid potential sources of evidence that her desired state of affairs does not obtain, seek the company of people likely to reassure her

---

26 Sartre suggests that engaging in self-deception is not the product of “a reflective, voluntary decision, but of a spontaneous determination of our being. One puts oneself in bad faith as one goes to sleep and one is in bad faith as one dreams.” (pg. 182)
that her false belief is accurate, avoid interacting with people who are likely to challenge the accuracy of her false beliefs, and so on.

On the other hand, suppose someone adopts an attitude of diligent resistance to the acquisition of false beliefs. In that case, she will try to avoid behaviors that make self-deception more likely to occur, or easier to sustain. For example, rather than talking to people who are likely to encourage her self-deception, she might spend time with people who are likely to give her accurate descriptions of her circumstances, honest evaluations of her character, and the like. Furthermore, she might search for evidence relevant to the truth or falsity of her beliefs in respected books or articles, or she might gather evidence by consulting people with relevant expertise. Finally, she might remain watchful for circumstances in which her anxious desires are likely to motivate her to adopt a false or unjustified belief. I argue that fulfilling a Kantian duty to aim diligently at truth in the formation of our judgments consists in engaging in these and similar behaviors.

Some might object that, even if most instances of self-deception could have been avoided if the agent had not adopted an attitude of acquiescence, adopting such an attitude does not embody a maxim of any sort. In that case, even if we accept my suggestions regarding the connection between acquiescence and adopting false beliefs, most cases of engaging in self-deception remain beyond the pale of Kantian permissibility and impermissibility. This objection may seem legitimate given that few, if any, self-deceivers explicitly state the principles on which their acquiescence depends. However, as I stated earlier, someone need not formulate a maxim explicitly, at any time, in order to act on it. Drawing an analogy between acquiescing in the formation of false beliefs and failing to give aid to people in need may be helpful. We can imagine a
case in which someone consistently fails to offer assistance to people who need it. He never goes out of his way to look for people who need assistance, and on the rare occasions when he notices that someone could use his help, he gets easily distracted and typically forgets that anyone could use his assistance. Even if such a person does not explicitly state his intention not to benefit others, his behavior embodies a principle something like “I will not make an effort to help people in need.” Similarly, we can plausibly claim that most self-deceivers act in ways that embody something like the principle of acquiescence, even if they do not state their intentions to allow themselves to adopt false beliefs.

Now the most urgent task for the Kantian is to show that a maxim of acquiescence would fail the categorical imperative test. Kant famously identifies two ways in which a maxim can fail the categorical imperative test. First, “some actions are so constituted that their maxims cannot without contradiction even be thought of as a universal law of nature, much less be willed as what should become one.” Second, “in the case of others, this internal impossibility is not to be found, but there is no possibility of willing that their maxim should be raised to the universality of a law of nature, because such a will would contradict itself.” The claim that a maxim of acquiescence fails in the former way is implausible. There does not seem to be a contradiction involved in imagining a world in which people allow themselves to believe whatever pops into their heads, whether or not the beliefs are rationally justified.

Nevertheless, I argue that someone cannot rationally will that a maxim allowing us to refrain from working diligently to avoid false judgments become a universal law of

---

27 Kant Ak. 424
28 Kant Ak. 424
nature. Of course, taking a moment to imagine what the world would be like if people did not exercise due care to ensure that their beliefs were accurate reveals that most of us would not want such a world to come about. Chaos would almost certainly ensue if people’s judgments regarding their environments, the consequences of their behavior, and even their own intentions were constrained solely by their desires rather than a diligent effort to arrive at true beliefs. However, merely showing that most (or even all) people happen to want to avoid a scenario in which no one works to ensure that their beliefs are true is not sufficient to show that the maxim of acquiescence fails the categorical imperative test. Basing the permissibility of acts that embody a given maxim on the contingent desirability of the results of acting on that maxim brings into ethics the heteronomy that Kant so vehemently warned against. Any effort to show that acting on the maxim of acquiescence is contrary to duty must remain consistent with Kant’s view that duty makes its demands independently of our contingent preferences and desires.

Taking a closer look at what a maxim of acquiescence involves, and borrowing some helpful considerations from Barbara Herman, we can show that willing such a principle into universal law is inconsistent with some of the fixed constraints on rational willing. In “Mutual Aid and Respect for Persons,” Herman rightly stresses that in

---

29 In *Groundwork II* Kant argues that “If the will . . . goes outside itself and seeks the law that is to determine it in the character of any of its objects, then heteronomy always results. The will in that case does not give itself the law, but the object does so because of its relation to the will. This relation . . . admits only of hypothetical imperatives: I ought to do something because I will something else. On the other hand the moral, and hence categorical, imperative says that I ought to act this or that way even though I did not will something else. (441)

willing some particular end, we are not merely wishing that a given situation would come about. Rather, reason demands that, insofar as we will the end, we take on a further commitment to bring about whatever means are required, as well as some sufficient means, for bringing about the desired end. She goes on to identify three types of means that may be available to secure our particular ends: (i) our own skills and capacities (ii) animate and inanimate objects and (iii) other people, who can willingly use their abilities to realize our ends. Furthermore, there are some ends (for instance, the continued ability to exercise our rational agency) that simply must be realized if we are to function as rational agents who adopt and pursue various ends.31 Insofar as we will anything at all, we ought rationally to will these necessary aims, which Kant characterizes as the conditions of our “power to set an end.”32 So we can claim that (i) we ought rationally to will whatever means are required to secure our ends and that (ii) there are ends that, insofar as we exercise our rational agency at all, we are committed to willing. Given this background, we can identify important ways in which a maxim of acquiescence is inconsistent with certain rational demands on willing.

The inconsistencies arise because actions that embody the principle of acquiescence thereby embody an indifference to the truth of the agent’s judgments. Elevating this indifference to truth to a universal law would radically undermine our efficacy as agents, and as a consequence, it would rob us of the first and third of the three types of means for securing our ends that I described above. We cannot either guarantee or justifiably assume that the ends we will pursue in the future can be secured

32 Immanuel Kant Doctrine of Virtue, Part II of Metaphysics of Morals
without the use of our own agency or the agency of others. So willfully undermining the efficacy of all rational agents is strikingly inconsistent with the pursuit of our ends.

To help illustrate the connection between our aiming at truth in judgments and our ability to function as rational agents we can imagine the *Land of No Error*, in which, no matter how sloppily people pursue their various projects, some powerful and benevolent deity ensures that they are carried out successfully. If someone, as a result of irresponsible evidence-gathering, adopts the false belief that a glass of poison is cool, refreshing water and decides to drink in order to relieve her thirst, the deity counteracts the toxin’s harmful effects and kindly quenches her thirst as well. Or, suppose that someone trying to rescue a child drowning in a nearby pool pauses at the water’s edge to eat a sandwich. In that case the deity will once again intervene and save the child’s life. Our world, however, is clearly *not* a “land of no error,” and in the actual world, our efficacy as agents depends, largely, on our making some effort to arrive at accurate judgments regarding our circumstances, the effects of our actions, and the nature of our goals. In the absence of some external force that faithfully sets things aright when our projects teeter on the brink of failure, the very possibility of our successfully functioning as rational agents often depends on our taking care to make accurate judgments. So, willing into universal law a maxim of acquiescence leads to inconsistency because such a law would severely undermine both our efficacy in securing our particular ends (including those ends that we are rationally committed to pursuing) and the efficacy of those who strive to help us realize our aims.

Furthermore, if we examine what is involved in the violation of a duty to guard against the acquisition of false judgments, what we find fits nicely with Kant’s characterization of the violation of duties in the *Grounding*. He argues that
If we attend to ourselves in any transgression of duty, we find that we actually do not will that our maxim should become a universal law – because this is impossible for us – but rather the opposite of this maxim should remain a law universally. We only take the liberty of making an exception to the law for ourselves (or just for this one time) to the advantage of our inclination. (424)

Following Kant’s recommendation, if someone “attends to herself” when she practices self-deception, she will recognize that she does not will that all people refrain from rigorously aiming at truth in the formation of their beliefs. In fact, she does not even will that she alone refrain from aiming at truth in her judgments all the time. Rather, she hopes to make her present case an exception to a general policy, which purports to hold for everyone all the time, of forming only those judgments that we have good reason to form – or, in certain circumstances, forming only those judgments that we have no good reason to refrain from forming.33

Some might object that I simply have not addressed the problem of self-deception that I aimed to resolve. After all, this section purports to address the worry that some people behave in seemingly immoral ways by acting on beliefs they acquired through irresponsible epistemic practices and that Kant attributes permissibility and even moral worth to their actions. After my discussion and defense of the duty to aim at truth in our judgments, I still have to accept the, perhaps implausible, claim that Boyce’s irresponsible administration of painful, ineffective treatments to his patients was both morally permissible and morally worthy.

I think the Kantian has a fairly plausible response to this worry. However, our estimation of the force of the objection and the adequacy of the response will depend on

33 Of course, as I mentioned previously (see note 24) this general policy to which the self-deceiver hopes to make her case an exception is not so strict that it rules out all cases in which someone thinks
how seriously we take the claim that people believe those claims that they accept through self-deception. To illustrate my point we can consider a case in which Roger develops, through self-deception, the clearly false belief that he is an extremely promising writer. Although he has not been able to publish any books, and the few stories that he published met with lukewarm reviews, he comes to believe that few critics respond favorably to his work because it is so far ahead of its time. Furthermore, he bolsters his belief that he has great, but under-appreciated talent by focusing on his close friends and family members’ assurances that he is a good writer, even though neither his friends nor his family members have any literary expertise. Perhaps Roger finds this belief somewhat difficult to sustain, and there are moments when he wonders whether he really has great, but unrecognized skill. However, when he accepts, through self-deception, the idea that he has considerable talent, this belief does not (at least not to him) appear different in kind from his other (justified) beliefs. In other words, although he may hold some beliefs to a greater degree of certainty and others to a lesser degree, his belief that he is a great writer does not in any way bear the mark of a false belief when he is self-deceived.

Now, whenever someone behaves deliberately, the more specific actions of which his behavior consists are heavily influenced by a background made up largely of the agent’s beliefs about his own capacities, features of his environment, the effects of his actions, the needs and intentions of other people, and so on. Furthermore, any number of the particular beliefs that help form the background of someone’s actions may be beliefs that he agent accepts through self-deception, and these beliefs will not bear any
distinctive markings of their falsity. The fact that those beliefs that someone holds through self-deception are indistinguishable (from the agent’s perspective) from those that are held justifiably makes room for the claim that Boyce’s actions are morally worthy. We can plausibly claim that an agent does not deserve any more blame for drawing on those beliefs that he sustains through self-deception when he chooses a course of action than he deserves for drawing on those beliefs that he is justified in holding (whether they are true or false). Rather, if self-deceived agents deserve blame at all, they should be blamed for irresponsibly allowing false judgments to enter (and remain within) the web of background beliefs that help determine their behavior.

Although I think the Kantian’s account of the attribution of moral worth to self-deceived agents seems plausible, her account of the attribution of permissibility may seem more problematic. After all, the claim that Boyce’s subjecting his patients to a painful and marginally effective therapy is permissible remains counterintuitive. Nevertheless, I think the Kantian’s claim that Boyce’s administration of the ineffective treatment is, by itself, permissible is more plausible than it may appear initially. We can best defend the approach I described by drawing an analogy between the Kantian’s treatment of Boyce’s case and his theory’s treatment of the following case. Suppose a doctor, whom I’ll call Frank, consults the best scientific views of his day and acquires the justified, but false, belief that some painful and ineffective treatment is really a cure for some ailment. Acting on the universalizable principle, “Use your talents to aid people in need,” Frank gives the dangerous treatment to his patients.

Insofar as we accept Kant’s view that willing in accord with certain rational constraints constitutes the moral law, we will accept the claim that Frank’s administering the treatment is permissible. Accepting this latter claim makes room for
the view that Kant’s attribution of permissibility to Boyce’s actions is plausible. Both Frank and Boyce act on the same universalizable maxim, and as I stressed above, the beliefs that an agent comes to accept through self-deception may be indistinguishable, from the agent’s point of view, from beliefs acquired in an epistemically justified way. Given these considerations, I think the claim that Frank acted permissibly is no more plausible than the claim that Boyce’s administering the ineffective treatments was, by itself, permissible.34

Of course, these considerations merely suggest that if Frank’s actions are permissible, we have some reason to judge that Boyce’s providing the ineffective therapy was also permissible. So someone who claims that neither Boyce nor Frank’s administration of ineffective medical treatments is permissible might nevertheless reject my approach. I think such a claim merely expresses a general dissatisfaction with the way autonomous ethical views determine the rightness and wrongness of actions, rather than particular worry about the adequacy of Kant’s theory for dealing with self-deception. In other words, those who claim that both Frank and Boyce acted wrongly when they administered the treatments will certainly be dissatisfied with the Kantian response to the problem of self-deception that I have outlined. However, this dissatisfaction is merely a product of more fundamental doubts about the plausibility of

34 If we consider the fact that Boyce acquired the belief on which he acts through epistemically irresponsible practices while Frank was justified in maintaining the belief on which he acted, then many of us begin to feel intuitively that Boyce’s behavior is, in some sense, more immoral than Frank’s. Furthermore, I suspect that even those who believe that both Frank and Boyce acted impermissibly will judge that, given this information about the agents’ belief acquisition, Boyce’s behavior is, in some respect, more immoral than Frank’s. Clearly the approach I am offering on Kant’s behalf can accommodate this intuition. On the view I am describing, Boyce’s behavior is more immoral because it runs contrary to the duty to rigorously aim at truth in one’s judgments.
Kant’s method of attributing rightness, wrongness and permissibility to actions, and these more fundamental doubts need not involve any mention of the problem of self-deception.

I’ll address one further worry before I conclude. I have argued that in cases like Boyce’s, in which a self-deceived person acts on permissible principles and worthy motives and nevertheless commits seemingly immoral actions, the unworthy aspect of the self-deceiver’s behavior is her failure fulfill her duty to aim at truth in her judgments. It seems plausible to suggest that in many cases of self-deception, perhaps in the vast majority of cases, the agent does fail to carry out this duty. However, I have left open the possibility that in some cases, people may go to extraordinary lengths to ensure that their beliefs are accurate (perhaps going well beyond what a duty to aim at truth demands) and nevertheless accept a false belief through self-deception. Such people may draw on the judgments that they form through self-deception and engage in apparently immoral behavior that embodies a permissible maxim and is performed from the motive of duty, and my strategy cannot avoid the conclusion that their actions are morally permissible in every respect.

Such cases are certainly possible, and I concede that when they arise I cannot claim that the self-deceived agent acts impermissibly. However, I do not think that this possibility poses a problem for the approach I have been developing. As I stated earlier, the fundamental problem of self-deception for Kant’s ethics is that an agent who acquires false beliefs through epistemically irresponsible practices and who wills according to certain rational constraints can perform apparently seriously immoral acts that are, on Kant’s view, both permissible and morally worthy. Admittedly, some cases might arise in which agents who carry out their duty to aim at truth nevertheless acquire
false beliefs through self-deception and, acting on these false beliefs, behave in seemingly immoral ways. However, by fulfilling this duty, that is, by taking due care to ensure that their beliefs are accurate, such agents avoid the charge of epistemic irresponsibility. So we can plausibly treat such cases the way we treat cases in which agents adopt and act on justified, but false, beliefs. In either case, if the agent’s actions embody a permissible principle and he acts from the motive of duty, then no matter how objectionable his actions appear, there simply aren’t any features of his behavior that ought to be described as immoral or unworthy.

VI

Some interesting results have emerged from this discussion, and it will be worthwhile to offer a brief overview of the more important ones. To begin with, I tried to show, as Darwall has pointed out and Kant himself seemed to recognize, that self-deception poses some particularly serious problems for Kant’s ethics and other autonomous ethical theories. More precisely, Kant’s view seems committed to attributing moral permissibility and moral worth to the seemingly immoral behavior of agents who act on false beliefs that they acquire through epistemically irresponsible practices. Also, such agents’ false beliefs might render them unable to use Kant’s theory to accurately determine what they ought morally to do. I argued that, perhaps contrary to initial appearances, Kant’s view has the resources to respond fairly plausibly to each of these problems. I further suggested that, in order to respond to two of the four problems of self-deception that I described, the Kantian has to claim that we have a duty to aim diligently at truth in the acquisition of our beliefs. So, on Kant’s view, and perhaps a number of other autonomous ethical theories, epistemic irresponsibility is, in part, a species of moral irresponsibility.