‘Early Stage’ Instrumental Irrationality: Lessons from Apathy

Annemarie Kalis, Stefan Kaiser

Philosophy, Psychiatry, & Psychology, Volume 25, Number 1, March 2018, pp. E-1-E-12 (Article)

Published by Johns Hopkins University Press
DOI: *https://doi.org/10.1353/ppp.2018.0000*

'Early Stage' Instrumental Irrationality: Lessons from Apathy

Annemarie Kalis &
Stefan Kaiser

Abstract: Understanding different forms of irrationality is an important aim in both philosophy and psychology, and the relation between everyday and pathological irrationality is a recurrent theme in the philosophy of psychiatry. Most work on irrationality in these different disciplines focuses on situations where the agent reaches a conclusion regarding what to do, but then somehow does not act on this conclusion. In this article, we argue that this is not the only form instrumental irrationality can take. The article attempts to broaden the perspective on instrumental irrationality by analyzing situations where an agent can be called instrumentally irrational for failing to reach a conclusion regarding what to do. We discuss two possible 'early stage' problems that might explain this kind of irrationality: lack of clarity about one's goals, and problems in determining possible means to attain one's goals. In philosophy, early stage irrationality is sometimes discussed under the heading of accidie, which is linked to depression. We try to show that this analysis cannot provide a substantial account of early stage irrationality. Instead, we argue that accidie is closer to the psychiatric symptom of apathy, and we explore how recent insights in apathy provide a fruitful basis for deeper understanding of early stage instrumental irrationality. We conclude by showing that these insights also shed new light on the capacities required for being instrumentally rational.

Keywords: *Accidie,* decision making, psychopathology, weakness of will, rationality, goals

As we all know, people often do not do what would be the rational thing to do. Both psychologists and philosophers have long been interested in explaining this aspect of the human condition. Also, the relation between everyday irrationality and pathological breakdowns of rationality is a familiar topic of discussion in psychiatry. It is not merely the failures themselves that present interesting questions; there is also the hope that, by understanding when and why we violate rational norms, we might get a firmer grasp on what it means to *meet* such norms, and thus gain a deeper understanding of the rational capacities of human beings. The capacity we focus on herein is our capacity for practical rationality, understood as our ability to guide our actions on the basis of rational norms. The norms that will be at stake here are norms of *instrumental* rationality or means-end rationality, defined in terms of coherence between one's goals and the means one adopts to achieve those goals (Kolodny & Brunero, 2016). Whether or not someone is violating norms of instrumental rationality, thus, depends on a person's goals; whereas taking the slow train to work might be instrumentally irrational for someone who wants to get to work fast, it might be perfectly instrumentally rational for someone with a different goal (such as meet-

ing a friend who is traveling on the slow train).[1] The observation that forms our starting point is that the existing literature in both philosophy and psychology focuses almost exclusively on certain forms of instrumental irrationality, while ignoring certain other interesting forms. Our aim is to develop a preliminary analysis of some of those other kinds of instrumental irrationality, inspired by recent psychiatric insights in the phenomenon of apathy.

The kind of irrationality that is usually investigated in both philosophy and psychology is the type of case in which the agent knows what her goals are and how to realize them, but somehow fails to translate these insights into action. The mismatch is, thus, located between the agent's conclusion of means-end reasoning and her behavior. In philosophy, the study of such mismatches has traditionally focused on the phenomenon of *akrasia* or weakness of will (Davidson, 1969). Weakness of will is usually defined as intentionally going against one's own judgment regarding what would be best to do (Davidson, 1969). Naturally, agents have multiple, and often conflicting, desires, and goals, and the main point of decision making is to judge in specific situations which of these goals to act on. In cases of akrasia, an agent reaches such a judgment, but nevertheless proceeds to act on a different goal or desire. An important philosophical question regarding akrasia is whether it is conceptually possible to act contrary to one's own best judgment. If not, akrasia is not a class of *actions*, but of mere (non-intentional) behavior—a distinction that would have implications for the possibility to ascribe responsibility. A second important question is how such actions or behaviors can be explained psychologically. In one sense, akratic behavior is perfectly intelligible; after all, the agent is aiming at the realization at one of his or her goals (Davidson, 1990). However, what *does* require explanation is why the agent did not act on the goal or desire that he or she judged to be most important or valuable; here, philosophers have made ample use of a variety of psychological insights (Davidson, 1969; May & Holton, 2012). Recently, Richard Holton has shifted the course of the debate by arguing that weakness of will should not be understood as going against your best judgment, but as revising your intention without warrant, a description that suggests the involvement of different psychological processes (May & Holton, 2012).

The observation relevant for our aim is that, whether one adopts a traditional or a 'Holtonian' understanding of weakness of will, in both cases the implicit assumption is that there is nothing wrong with the way the agent reasons toward a conclusion regarding what to do, nor with the conclusion itself; the problem is that the agent either forms the 'wrong' intention, or forms the right intention but does not act on it. In psychology, the situation is similar; here, instrumental irrationality is usually discussed under the heading of self-control failure, often defined as the failure to resist temptation. Most research on self-control failure focuses on the 'late-stage failure' of not following up on one's decision under the influence of temptation; although the rational thing to do usually is to act in accordance with our general life goals, in the heat of the moment we are prone to succumb to the temptation of immediate rewards, leading to irrational behavior (Baumeister, Vohs, & Tice, 2007). Research has, for example, focused on the difficulties people experience in keeping their resolutions or sticking to their diets (Massey & Hill, 2012). Another relevant area of psychological inquiry focuses on discordances between explicit and implicit measures of preferences, and their differential impact on choice (Hofmann, Friese, & Strack, 2009). Under certain conditions we *say* we prefer one thing (a statement that is often interpreted as a conclusion of explicit reasoning), but nevertheless show contradictory choice behavior, which can be predicted with implicit measures. This raises the question of how to understand this discordance between what we say we prefer and what we actually choose.

Most of the work on instrumental irrationality, thus, focuses on mismatches between, on the one hand, the agent's conclusion regarding what to do, and her actual behavior. Our main claim is that this is an unnecessarily restricted interpretation of what it means for an agent to be instrumentally irrational. After all, instrumental irrationality is generally defined as a mismatch between the agents' goals and his or her behavior.

Now although such a mismatch between goals and behavior *might* present itself as an agent reaching a conclusion but not acting on it, our point is that it does not *need* to present itself as such. Instead, we argue that a mismatch between goals and behavior also often manifests itself as agents *not reaching a conclusion*. Agents can violate norms of instrumental rationality in failing to become aware of their relevant goals in concrete situations, or in failing to determine how to realize their goals. This then results in the agent not reaching any kind of conclusion regarding the question what to do.

The article is structured as follows. First, we introduce a set of everyday examples showing that agents often seem to have trouble determining relevant goals in a specific situation, or have trouble determining how to realize these goals. We argue that these phenomena are forms of instrumental irrationality, or, in other words, mismatches between the agent's ends and the means she adopts. Now, although such cases are not regularly discussed in the literature on practical irrationality, they are now and then mentioned in a specific theoretical context under the heading of *accidie*. The next section provides an overview of this philosophical discussion, and show that, although it provides some promising leads, it falls short as an account of the kind of examples under investigation. Next, we argue that the psychiatric symptom of apathy might shed interesting light on these forms of instrumental irrationality. In the final two sections, we show how insights in apathy could contribute to the explanation of different forms of instrumental irrationality, and to the understanding of the capacity for instrumental rationality itself.

'Early Stage' Instrumental Irrationality

All of us are familiar with the experience of struggling to get some aspects of our lives, or maybe even our lives in general, on track. For example, we might be vaguely aware that we should try to find another job, or change something about our family relationships, or do something about our physical condition. Maybe we have an indeterminate feeling of failure and dissatisfaction, without having any clear ideas on how to change things. Or we might have some ideas: We should talk to our boss, go to the gym more often, or avoid doing the shopping while hungry. Nevertheless, days and weeks pass without us making the changes we consider to be the right ones. Somehow, we do not seem to find the 'right' moment for making concrete changes.

Our claim is that, in such situations, agents fail to reach their goals because they somehow do not manage to reach a conclusion regarding what to do. Before introducing two different ways in which this can happen, we should first address a basic concern regarding the possibility to understand such situations in terms of instrumental irrationality. The worry is that the phenomenon we somehow consider to be problematic is not a specific action, but *inaction*. But how can an individual moment of inaction be called instrumentally irrational? After all, we are not rationally required to realize all our aims all the time. This means that there is no specific point in time about which we can say that, *at that point in time*, we are not doing what we should do. We might still come to a conclusion regarding what to do tomorrow, and realize our aim next week, or next year. This problem does not arise in standard discussions on instrumental irrationality, precisely because these focus on discordances between conclusion and action *at a specific point in time*. Akrasia, for example, is defined as an agent drawing, at a certain point in time, a conclusion regarding what should be done right now, and directly afterwards doing something that blatantly contradicts that conclusion: The agent is thus literally 'caught in the act,' the irrationality is located then and there. In the cases we described, it is not clear exactly when and where the irrationality takes place, and this seems to preclude any ascription of such irrationality. However, this objection only gains traction on the assumption that only individual doings at certain points in time are suitable carriers of the label 'instrumentally irrational.' But why would this be so? After all, even in standard cases of weakness of will, it is not the action that is irrational: It is the *agent* who is not realizing her ends. The only suitable carrier of the label 'instrumentally irrational' is, therefore, the agent

herself. With regard to inaction and failure to reach a conclusion, we can, therefore, say that insofar as we have reason to ascribe goals to an agent, and the agent subsequently fails to realize these goals, notwithstanding sufficient ability and opportunity to do so, the agent can be called instrumentally irrational.

So, how could we gain further understanding of the problem of agents failing to reach a conclusion regarding what to do? Without wanting to claim that there is one definite structure to be found here, our proposal is that one can distinguish at least two problems that might lead to 'early stage' instrumental irrationality: a lack of clarity about one's goals in specific situations, and a lack of good ideas regarding how to achieve one's goals. In the remainder of this section, we argue why we think these problems could lead to a mismatch between an agent's goals and her behavior. What we hope to achieve by discussing such problems is to show that there is room in the concept of instrumental irrationality for the idea that something might already go awry *before* we reach any kind of conclusion on what would be the best thing to do.

To start with the first problem, often agents seem to be unclear regarding what their goals are. In such cases, goals do not become sufficiently 'active' in concrete situations and, therefore, fail to guide decision making. Now, whereas few philosophers or psychologists will want to deny that this is a familiar phenomenon, many will object that this problem cannot be understood as a form of instrumental irrationality. After all, if the agent is not clear about what her ends are, why should we presume that she nevertheless has certain specific ends? And if it is not warranted to ascribe certain ends to such an agent, we cannot call her instrumentally irrational for not realizing these ends. Maybe one could instead make a case for calling such an agent non-instrumentally irrational: maybe the problem is that the agent does not act in accordance with ends she *should* have? However, this would burden us with the weighty task of showing that agents are rationally required to have certain specific ends. We do not know whether such a case could be made, but for our purposes we do not need an answer to this question. Here, the plausibility of our analysis hinges on the question *what is required to ascribe a certain goal to an agent*. We want to answer this question by arguing for a dual claim: First, that an agent who is unclear about a goal might nevertheless have that goal (this is the ontological part of the claim), and second that, in cases where agents are unclear about their goals, there are certain epistemological clues we can use for finding out which goals to ascribe to such agents (the epistemological claim).

In the current literature, most philosophers and psychologists want to avoid reducing the notion of a goal to mere behavior (which would mean that agents only have a goal insofar as they are engaged in realizing it) or to something biological (meaning that agents only have a goal insofar as they are, for instance, in a certain kind of brain state). According to current psychological definitions, goals are considered to be "subjectively desirable states of affairs that the individual intends to attain through action" (Kruglanski & Kopetz, 2009, p. 29). The question thus becomes: What does it mean for an agent to intend to attain a goal state? We do not think there is an always a crystal clear, objectively valid answer to the question whether an agent has a goal or not; in some cases, it might just be an ambiguous matter. However, the literature does not provide any grounds for thinking that clear conscious awareness is a conceptual requirement for goal possession; in fact, several lines of research have argued that goals and intentions can operate outside conscious awareness (Kruglanski & Kopetz, 2009).

This brings us to the epistemological question: How do we determine whether a person has a certain goal or not? Imagine an agent whom we observe to be always unhappy when coming home from work, and who regularly finds excuses to stay home from work. Someone might also say things like, 'I will be stuck here for the rest of my life,' or 'I really hate my job.' Here it is underdetermined whether this person has a goal of wanting to find another job, or not. However, the fact that the agent does not spontaneously report that she has (or does not have) this goal, does not need to be the end of the story. We can use the clues provided by an agent's doings and sayings to actively ask such an agent what she actually wants to do. In such a

conversation, the agent might respond to our questions in such a way that both we, and the agent herself, gain more clarity on the question whether or not she wants to find another job. Of course, to count as sufficient grounds for ascribing a goal, the person's response to our questions should be more 'substantial' than just checking a box after the question, 'Do you want to find another job?' This shows that there remain important questions to be asked about the conditions under which we can legitimately believe what an agent tells us (when is her response 'substantial enough'?). What we have merely wanted to show here is that, in different kinds of situations, an agent's doings and sayings provide worthwhile starting points for examining which goals to ascribe to an agent in cases in which the agent himself or herself is unclear about this. And, insofar as we do have indications to ascribe specific goals, an agent not reaching a conclusion regarding how to realize such goals can be said to be instrumentally irrational.

So far, we have explored the possibility that an agent might have a certain goal (according to the criteria discussed), but is not clear about, or attentive to, that goal for taking the steps that are necessary to realize it. This is the first type of early stage instrumental irrationality. The second type we want to distinguish is displayed in cases where an agent is clear about her goals, but fails to determine the steps necessary to realize them. People might want to change their relationships, but do not get to the point of developing concrete ideas on how to achieve that end. Or, they do generate ideas, but only ideas that they themselves consider to be inadequate ('we might go into family therapy—oh no, that would be a disaster'). What makes such cases forms of 'early stage' irrationality is that, due owing to a lack of adequate ideas, the process of decision making cannot continue, and no conclusion will be reached. In contrast, in traditional examples of instrumental irrationality, at least some conclusion *is* being reached, even though the agent subsequently does not act on that conclusion. Cases in which an agent fails to determine adequate steps for realizing his or her goals do not require any difficult theoretical moves to understand them as a kind of instrumental or means-end irrationality; because the agent does not manage to determine adequate means for realizing her aim, she does not adopt the means that are necessary for realizing those aims. We, therefore, are brief with regard to this type of problem, and move on to see what has been said about these kinds of early stage instrumental irrationality in the literature so far.

Accidie

Although the kind of cases introduced in the previous section are not regularly discussed in the literature on practical irrationality, the basic problem of agents not reaching a conclusion regarding what to do has been mentioned in certain philosophical debates under the heading of *accidie*. In the Middle Ages, the term *accidie* (sometimes also called *acedia*) meant to describe a lack of caring and spiritual sorrow that was sometimes found in monks. In contemporary philosophy, the phenomenon regained attention by Michael Stocker's influential paper 'Desiring the bad' (1979), where he mentions the phenomenon and links it to related experiences:

> Through spiritual or physical tiredness, through accidie, through weakness of body, through illness, through general apathy, through despair, through inability to concentrate, through a feeling of uselessness or futility, and so on, one may feel less and less motivated to seek what is good. One's lessened desire, need not signal, much less be the product of, the fact that, or one's belief that, there is less good to be obtained or produced. (Stocker, 1979, p. 744)

For Stocker, the existence of phenomena like *accidie* supported his claim that 'the (believed) good need not attract.' In his view, someone suffering from *accidie* still sees things as good, and thus seems to have reasons for action, but is not motivated to act on those reasons.

This has remained the main focus of philosophical discussions on *accidie*; the question usually addressed is whether this phenomenon presents an objection to motivational internalism about reasons. Motivational internalism covers the idea that someone can only be said to have a reason if he or she is at least to some extent motivated to act accordingly. Accidie presents a problem to this view in that it seems to be a case where

someone can accept something as a reason for action ('Yes, I know I should find another job!') without the agent being motivated to act on that reason. Most participants in this debate presuppose that accidie concerns a complete *absence* of motivation; it focuses on the conceptual question whether it is possible to have reasons for action and nevertheless not be motivated by them. Moreover, most authors focus on *moral* reasons and (lack of) corresponding moral motivation, although it is usually assumed that the problem is ultimately about the relation between reasons and motivation in general.

In the philosophical literature, accidie is often discussed as a symptom of depression, or even as being equivalent to depression (Roberts, 2001). This link between accidie and depression is most explicitly made by Solomon (2014) in his autobiographical account of depression, when he argues that in the Middle Ages, the term *accidie* "seems to have been used almost as broadly as the word depression in modern times, and it described symptoms familiar to anyone who has seen or felt depression" (p. 293).

Although the notion of accidie raises interesting questions, we think that so far it falls short of an analysis of the kind of early stage cases of instrumental irrationality introduced above. First, discussions on accidie often suggest that what is at stake is a *complete lack* of motivation. This is because accidie is brought forward as a *conceptual possibility,* which as such offers a counterargument against motivational internalism (Cholbi, 2011). However, as Cholbi himself concludes, most authors clearly state that they also consider accidie to be a real-life phenomenon that frequently occurs in the world. But how should we make sense of the idea of *complete absence of motivation?* What conception of motivation is used here? The literature on accidie does not provide any answer to these questions. The second shortcoming is that several authors in this debate use somewhat 'lazy hand waving' to psychiatry. They mostly just briefly refer to clinical descriptions of psychiatric phenomena, such as apathy or depression (Roberts, 2001). However, these references are usually not supported by any actual arguments or empirical information. One misleading effect of such hand waving is described by Cholbi (2011), who argues that even if it might make sense to claim that depression is associated with a decrease in motivation, it is certainly *not* associated with a decrease in motivation to obey moral norms, which is the focus of most philosophers writing about *accidie*. This brings us to the third shortcoming: The discussion of *accidie* is restricted to its being a counterargument to motivational internalism, primarily concerning moral reasons. It is not discussed as a type of instrumental irrationality that deserves analysis in itself. This is not exclusively a problem of the philosophical literature; in the psychological literature on self-regulation failure, the forms of early stage instrumental irrationality discussed herein are hardly discussed either.

So, how should we proceed to develop a more satisfactory and illuminating account of early stage instrumental irrationality? Our suggestion is that fruitful psychological explanations might be found by taking a closer look at the psychiatric symptom of apathy, which description actually matches the philosophical definition of *accidie* much better than depression. Given that apathy in psychiatry refers to a reduction in goal-directed activity, even though apathetic patients can still be said to have certain general goals, this symptom seems to bear interesting similarities to the cases of early stage instrumental irrationality we focus on here. In recent years, we have examined certain mechanisms underlying apathy from the perspective of decision-making research, and we think this work could shed an interesting light both on *accidie* as a form of instrumental irrationality, and on instrumental rationality in general.

Apathy

As a first step, we should say a few words on why it would be a good idea to look at psychiatric symptomatology to acquire insight in everyday cases of instrumental irrationality. After all, one might hold that certain behavioral phenomena are labeled 'pathological' precisely because they follow an order that is categorically different from the order present in everyday behavior. However, this no longer seems to be the majority view. During the last 20 years, philosophers have shown increasing interest in the question how

'everyday' forms of irrationality relate to more substantial breakdowns of rationality in certain forms of psychopathology (Campbell, 2000). This increasing interest is paralleled in research on psychopathology. Traditionally, the symptoms associated with psychiatric disorders have been considered to be categorically different from "normal" human experience and behavior. This is contradicted by consistent findings that all psychopathological symptoms can also be observed in persons not fulfilling the criteria for a psychiatric disorder, albeit often with lower intensity or frequency (Johns & Van Os, 2001; Kaiser, Heekeren, & Simon, 2011). These observations have been accommodated in dimensional conceptions of psychopathology, which suppose a continuous distribution of symptoms across the population (for an overview of discussions on categorical and dimensional approaches during the development of the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders*, see Widiger & Sankis, 2000).

In psychiatry, a breach in the connection between reasons (or values) and motivation is usually not discussed in terms of depression, but in terms of apathy. In talking about apathy, we refer to the psychiatric use of the term. This needs to be mentioned because the term apathy is also now and then used in philosophical discourse, usually to describe a laudable state (first discussed by the Stoics) in which one does not have any desires. In psychiatry and neurology, apathy has been defined as an impairment of motivation or a quantitative reduction in goal-directed behavior (Levy& Dubois, 2006; Marin, 1990). We return to these definitions elsewhere in this article. Apathy can occur as a consequence of neurologic disorders as Parkinson's disease (Dujardin et al., 2007) and Alzheimer's disease (Robert et al., 2009). In psychiatry, apathy has been defined as a core negative symptom of schizophrenia (Kirkpatrick, 2014). Importantly, apathy can also occur in the context of depression, but depression is a broader syndrome associated with additional symptoms affecting mood, cognition and physiology.

Just like *accidie*, apathy has traditionally been seen as a motivational problem (Marin, 1990). However, the psychiatric definition of apathy as a motivational problem has recently been criticized by Levy and Dubois (2006, p. 916), who state that "it is unlikely that the concept of 'lack of motivation' represents the underlying mechanism responsible for apathy because it is a projective psychological interpretation of a given behavioral state." This criticism shows an ambiguity that is inherent in the use of the term motivation in psychiatry. Motivation has both been used to designate "a mental construct energizing action" (Shah & Gardner, 2008) and at the same time as a more descriptive term referring to the "behavioral, cognitive, and emotional *concomitants of goal-directed* behavior" (Marin, Biedrzycki, & Firinciogullari, 1991). The critique by Levy and Dubois is aimed mainly at the use of the term motivation as a description of the mechanisms underlying apathy and they consequently propose to define apathy not in terms of the underlying mechanisms but as an observable reduction in goal-directed behavior. Other authors would keep the term motivation in a descriptive sense, but similarly avoid using it as an explanatory mechanism.

If motivation is not used as an explanatory concept, the question arises of how apathy is explained in psychiatry. It is common to distinguish different aspects of apathy that are thought to correspond with different underlying processes or mechanisms, aspects that together constitute the observed lack or decrease of motivation or goal-directed behavior that is observed. First, a key role has been attributed to dysfunctions of reward processing. Patients suffering from apathy in the context of neurologic and psychiatric disorders show reduced anticipatory pleasure, that is, they feel less pleasure in relation to an expected positive event (Kring & Barch, 2014). In other words, if the patient does not anticipate pleasure from outcomes, he will be less likely to engage in activities to arrive at these outcomes. Furthermore, learning from positive outcomes has been shown to be impaired in patients with apathy in several but not all studies (Gold et al., 2012; Hartmann-Riemer et al., 2017). If the agent's actions do not result in learning from positive outcomes, she will be less likely to engage in such activities in the future. Overall, it has been suggested that these rewards processing deficits might best be explained

by a general impairment in value representation.

Although this research has focused mainly on rewards processing, the integration of costs and benefits during decision making has recently been emphasized (Green, Horan, Barch, & Gold, 2015). This approach assumes that a decrease in goal-directed behavior is not merely the result of an impaired coding of rewards, but must more generally be explained by an overweighing of costs in relation to rewards. In other words, the reward is not worth the effort. Such a shift has been demonstrated in patients suffering from apathy in the context of psychiatric disorders (schizophrenia, depression) and neurologic disorders (Parkinson's disease; Chong et al., 2015; Hartmann, Hager, et al., 2015). The mechanisms underlying the impaired integration of costs and benefits include a reduced valuation of the reward at stake as well as a dysfunctional cost–benefit computation.

Another group of mechanisms refers to the cognitive operations required for the planning and implementation of actions. In neurologic disorders, a dysfunction of cognitive functioning in general and executive processes in particular has been shown to be associated with apathy (Andersson & Bergedalen, 2002). In psychiatric disorders, a relationship between cognitive function and apathy has also been observed, but seems to be less clearly delineated (Faerden et al., 2009). However, it has to be kept in mind that, in most studies, a broad range of cognitive functions was investigated, which might not all be equally relevant for planning and implementation of actions. More recently, we have suggested that pre-decisional cognitive processes might also be relevant for apathy (Hartmann, Kluge, et al., 2015). Specifically, the capacity to generate options for actions is negatively associated with apathy in patients with schizophrenia.

Overall, there is increasing evidence that both reward system and cognitive system dysfunction can contribute to apathy in patients with neuropsychiatric disorders. Reward system dysfunction concerns mainly an impaired coding of reward value, whereas cognitive dysfunctions are more diverse and include deficits in cost–benefit computation, planning, and option generation. It remains difficult to identify which reward and cognitive processes are dysfunctional in the individual patient. Even though most patients will probably show a combination of dysfunctional processes, a precise definition on the individual level would be of high importance for treatment. However, at the population level, current models of apathy include mechanisms of both reward and cognitive processing.

Explaining *Accidie*

Our suggestion is that these explanatory mechanisms underlying apathy in psychiatric patients could shed light on *accidie*, or early stage failures of instrumental rationality; they might provide valuable hypotheses regarding mechanisms that could explain why agents often seem to lack clarity on what their goals are, or how to realize them. At this point, it is important to emphasize that possible explanations of early stage irrationality do not explain *why apathy and accidie are irrational*. As mentioned, 'rational' and 'irrational' are normative predicates that apply to *agents*, and not to behaviors, processes, or psychological mechanisms. By calling human agents *rational* agents, we mean that they have *rational capacities*. In this sense, the predicate 'rational' applies to human agents *regardless* of what they are doing or not doing. But we also make distinctions in that we consider things the agent does, or does *not* do, as either manifestation of these rational capacities (e.g., cases where the agent adopts means that are suitable for attaining his or her ends), as 'misfirings' of these capacities (such as the different types of instrumental irrationality discussed here) or as neither manifestations nor misfirings (such as behaviors like sneezing). So, what we are explaining when we are pointing to certain psychological mechanisms involved in *apathy* or *accidie*, is how certain behavioral phenomena come about that we have already independently evaluated as forms of irrationality. Although this distinction does not help us to understand rationality or irrationality as such, it can provide important insights in how to change those behavioral phenomena that we consider to be irrational.

Regarding the mechanisms that might be relevant for explaining the two forms of early stage instrumental irrationality discussed, we can only

offer very preliminary hypotheses here, but we hope these can provide fruitful starting points for empirical investigation. To start with the problem of agents having a lack of clarity about their goals, work on apathy seems to offer some leads in that it points toward a role for, first, reward processing mechanisms, and second, cognitive dysfunction. These two putative mechanisms are also reflected in the definition of goals as 'internal representations of desired states,' that is, a goal requires a reward value to be linked to a cognitive goal representation. To start with the first: To start realizing a goal, an agent needs to be at least sometimes consciously aware of, and attentive to, that goal as a goal. This seems to require some kind of acknowledgment of the value of what the agent aims at. Being aware that one wants to find another job requires the agent to see the situation of having a better job as valuable and rewarding. In other words, difficulties in getting clear on what one's goals are might partially be difficulties in *seeing things as valuable and rewarding*. Representing something as a goal, thus, seems to require that the agent evaluatively connects with the representation of a certain end state. Second, and relatedly, problems with the cognitive aspects of goal representation could also impair the clarity and awareness of goals. In this context, different putative mechanisms have been invoked, such as the initial encoding of the goal representation, its maintenance, and its updating depending on the situational context (Barch & Dowd, 2010). However, the empirical evidence linking these specific cognitive aspects of goal representation to apathy is currently still limited. Regarding both cognitive representation and reward value of goals, it has to be kept in mind that experimental research has so far mostly focused on goals on a microscopic scale, for example, a monetary reward after a button press. Thus, it is not yet established whether these concepts and empirical findings translate to more complex and more temporally extended human goals.

So, what about agents who *are* aware of, and attentive to, their goals but who have difficulties in determining how to realize them? Here, the work on apathy seems to suggest that such difficulties might be explained by dysfunctions in the mechanisms underlying option generation.[2] The problem might be primarily a problem of quality (the agent generates just as many options as other agents, but the options generated are inadequate), or a problem of quantity that leads to a problem in quality (agents come up with so many options that it becomes impossible to compare and evaluate them, or they generate few options and, therefore, overlook the important ones). In everyday life cases, this might result in not realizing one's goals; in pathological cases, it might even result in a general reduction of goal-directed behavior (Levy & Dubois, 2006). Without options for action, no decision can be made, no intention formed, and no action taken.

So far, we have argued that phenomena of *accidie* or early stage irrationality, where agents do have goals but fail to reach a concrete conclusion on how to realize those goals, might be explained in terms of mechanisms that play a role in the psychiatric phenomenon of apathy. Additionally, we want to suggest that the mechanisms going awry in early stage instrumental irrationality, might also shed light on the explanation of certain more 'classical' types of instrumental irrationality. Here, we mean cases where the agent *does* reach a conclusion regarding what to do, but nevertheless does not act on that conclusion. Now whereas most explanatory hypotheses focus on the role of mechanisms related to action initiation or willpower (Baumeister et al., 2007), the insights developed so far suggest that the problem need not be a dysfunction in transforming conclusions into action. Another possibility is that the problem is *in the conclusion*: The conclusion might not relate in the right way to the agent's actual goals, or the conclusion might be based on a faulty assessment of the means required to realize those goals. This means that the mechanisms underlying apathy might also be able to contribute to the explanation of 'classic' cases of weakness of will or self-control failure.

A Wider View on Rational Capacities

This leaves us with the question: Could the insights developed so far also contribute to the

understanding of our capacity for instrumental rationality? After all, one of the reasons people investigate instrumental irrationality is that, by finding out why we often fail to act in accordance with our goals, we hope to get a clearer understanding of what it means to act in a way that *accords* with such goals. However, as stated, it would be a mistake to think that insights on mechanisms underlying apathy or other psychiatric phenomena might put us on the track of 'rational mechanisms' or of mechanisms that could explain what makes certain actions instrumentally rational.

We think that insight in possible ways in which agents might fail to reach a conclusion regarding what to do does provide information on the kind of capacities an agent needs to have for the predicate 'rational' to be applicable. In the beginning of the article, we stated that we would focus on *instrumental* rationality, a predicate that can be applied to agents insofar as they are capable of 'adopting the means required to attain one's goals.' The existing classical literature on weakness of will and self-control failure has taught us that being rational requires an agent to transform a conclusion regarding what means to adopt, into actual action. Authors differ on the question as to which capacities are needed to do this. Davidson, for example, thought that the crucial capacity was to transform an all things considered judgment into an unconditional judgment or an intention (Davidson, 1969). In contrast, contemporary authors such as Roy Baumeister and Richard Holton focus on the importance of willpower as a capacity (Baumeister et al., 2007; Holton, 2003). What we think our analysis could show is that there are other capacities required for being instrumentally rational.

First, an agent should be able to determine what his or her goals are. Against this idea, it has been argued that rational action might not require explicit deliberation (Arpaly, 2000) and, thus, might also not need to involve any conscious awareness of one's goals. However, although we certainly do not want to claim that goal-directed behavior requires constant or even frequent presence of one's goals in conscious awareness, we think that being able to act in a goal-directed way requires that an agent is able to represent a certain desired end state *as a goal*, and thus to be able to see that end state as something to be achieved. Practically, this implies that interventions aimed at strengthening people's rational capacities should focus not only on helping agents to transform their goals into concrete plans or intentions (Fishbach & Hofmann, 2015); it might be just as important to strengthen agents' capacity to determine what they really want, and to raise the awareness of their goals in relevant situations.

Second, being instrumentally rational requires not only the capacity to choose between different possible means, but also the capacity to come up with good ideas for realizing those aims. Option generation, which is not a specific mechanism, could be understood as an often-overlooked aspect of the capacity for instrumental rationality. It fulfills two (complementary) roles in instrumental rationality: It opens up an action space, and simultaneously narrows down that action space by relating one's possibilities to one's goals. After all, if we feel like it we could ponder an almost infinite number of ways to spend our holidays, or books to read. Right now, one could sum up an endless list of things one could do: Make more coffee, stand on one's head, but also book a cruise or throw all one's books through the window. And the fact that all these things are open for one to consider, the fact that one can determine for oneself how all these possible things to do could contribute to goal fulfillment, seems to be a crucial aspect of what it means for to be capable of goal-directed action (Kalis, Kaiser, & Mojzisch, 2013).

We have tried to do several things herein. Most important, we hope to have shown that instrumental irrationality can take a variety of forms, and that there is more to it than classical weakness of will or self-control failure. Second, we have tried to show that recent work on the psychiatric symptom of apathy could provide interesting hypotheses regarding the explanatory mechanisms underlying early stage instrumental irrationality. And finally, we have argued that a more thorough understanding of such early stage failures could shed light on the capacities agents need to be able to exercise instrumental rationality. We hope our suggestions could provide an impetus to both philosophical and the psychological discussions

on practical irrationality, and stimulate discussions on the relation between everyday irrationality and pathological dysfunctions.

Notes

1. In the philosophical literature, instrumental rationality is not the only conception of practical rationality being discussed. Many assume that certain actions can be called irrational, regardless of the agent's goals.

2. Although option generation is not a distinct psychological mechanism, work on option generation offers substantial leads concerning the mechanisms involved (Kaiser et al., 2013; Raab, de Oliveira, & Heinen, 2009).

References

Andersson, S., & Bergedalen, A.-M. (2002). Cognitive correlates of apathy in traumatic brain injury. *Cognitive and Behavioral Neurology*, *15*(3), 184–191.

Arpaly, N. (2000). On acting rationally against one's best judgment. *Ethics*, *110*(3), 488–513.

Barch, D. M., & Dowd, E. C. (2010). Goal representations and motivational drive in schizophrenia: The role of prefrontal–striatal interactions. *Schizophrenia Bulletin*, *36*(5), 919–934.

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, *16*(6), 351–355.

Campbell, P. G. (2000). Diagnosing agency. *Philosophy, Psychiatry, & Psychology*, *7*(2), 107–119.

Cholbi, M. (2011). Depression, listlessness, and moral motivation. *Ratio*, *24*(1), 28–45.

Chong, T. T.-J., Bonnelle, V., Manohar, S., Veromann, K.-R., Muhammed, K., Tofaris, G. K., … Husain, M. (2015). Dopamine enhances willingness to exert effort for reward in Parkinson's disease. *Cortex*, *69*, 40–46.

Davidson, D. (1969). How is weakness of the will possible? Available: http://philpapers.org.proxy.library.uu.nl/rec/DAVHIW.

Davidson, D. (1990). Paradoxes of irrationality, In: P. K. Moser (Ed.), *Rationality in action: Contemporary approaches* (pp 449–464). Cambridge: Cambridge University Press.

Dujardin, K., Sockeel, P., Devos, D., Delliaux, M., Krystkowiak, P., Destée, A., & Defebvre, L. (2007). Characteristics of apathy in Parkinson's disease. *Movement Disorders*, *22*(6), 778–784.

Faerden, A., Vaskinn, A., Finset, A., Agartz, I., Barrett, E. A., Friis, S., … Melle, I. (2009). Apathy is associated with executive functioning in first episode psychosis. *BMC Psychiatry*, *9*(1), 1.

Fishbach, A., & Hofmann, W. (2015). Nudging self-control: A smartphone intervention of temptation anticipation and goal resolution improves everyday goal progress. *Motivation Science*, *1*(3), 137.

Gold, J. M., Waltz, J. A., Matveeva, T. M., Kasanova, Z., Strauss, G. P., Herbener, E. S., … Frank, M. J. (2012). Negative symptoms and the failure to represent the expected reward value of actions: Behavioral and computational modeling evidence. *Archives of General Psychiatry*, *69*(2), 129–138.

Green, M. F., Horan, W. P., Barch, D. M., & Gold, J. M. (2015). Effort-based decision making: A novel approach for assessing motivation in schizophrenia. *Schizophrenia Bulletin*, *41*(5), 1035–1044.

Hartmann, M. N., Hager, O. M., Reimann, A. V., Chumbley, J. R., Kirschner, M., Seifritz, E., … Kaiser, S. (2015). Apathy but not diminished expression in schizophrenia is associated with discounting of monetary rewards by physical effort. *Schizophrenia Bulletin*, *41*(2), 503–512.

Hartmann, M. N., Kluge, A., Kalis, A., Mojzisch, A., Tobler, P. N., & Kaiser, S. (2015). Apathy in schizophrenia as a deficit in the generation of options for action. *Journal of Abnormal Psychology*, *124*(2), 309.

Hartmann-Riemer, M. N., Aschenbrenner, S., Bossert, M., Westermann, C., Seifritz, E., Tobler, P. N., … Kaiser, S. (2017). Deficits in reinforcement learning but no link to apathy in patients with schizophrenia. *Scientific Reports*, *7*, 40352.

Hofmann, W., Friese, M., & Strack, F. (2009). Impulse and self-control from a dual-systems perspective. *Perspectives on Psychological Science*, *4*(2), 162–176.

Holton, R. (2003). How is strength of will possible. In: S. Stroud & C. Tappolet (Eds.), *Weakness of Will and Practical Irrationality* (pp. 39–67). Oxford: Oxford University Press.

Johns, L. C., & Van Os, J. (2001). The continuity of psychotic experiences in the general population. *Clinical Psychology Review*, *21*(8), 1125–1141.

Kaiser, S., Heekeren, K., & Simon, J. J. (2011). The negative symptoms of schizophrenia: Category or continuum? *Psychopathology*, *44*(6), 345–353.

Kaiser, S., Simon, J. J., Kalis, A., Schweizer, S., Tobler, P. N., & Mojzisch, A. (2013). The cognitive and neural basis of option generation and subsequent choice. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 814–829.

Kalis, A., Kaiser, S., & Mojzisch, A. (2013). Why we should talk about option generation in decision-making research. *Frontiers in Psychology*, *4*(555), 10–3389.

Kirkpatrick, B. (2014). Developing concepts in negative symptoms: Primary vs secondary and apathy vs expression. *Journal of Clinical Psychiatry*, *75*, 3.

Kolodny, N., & Brunero, J. (2016, Spring). Instrumental rationality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/archives/spr2016/entries/rationality-instrumental/.

Kring, A. M., & Barch, D. M. (2014). The motivation and pleasure dimension of negative symptoms: Neural substrates and behavioral outputs. *European Neuropsychopharmacology*, *24*(5), 725–736.

Kruglanski, A. W., & Kopetz, C. (2009). What is so special (and nonspecial) about goals? A view from the cognitive perspective. In G. B. Moskowitz & H. Grant (Eds.), *The Psychology of Goals* (1st ed., pp. 27–55). New York: The Guilford Press.

Levy, R., & Dubois, B. (2006). Apathy and the functional anatomy of the prefrontal cortex–basal ganglia circuits. *Cerebral Cortex*, *16*(7), 916–928.

Marin, R. S. (1990). Differential diagnosis and classification of apathy. *American Journal of Psychiatry*, *147*(1), 22–30.

Marin, R. S., Biedrzycki, R. C., & Firinciogullari, S. (1991). Reliability and validity of the Apathy Evaluation Scale. *Psychiatry Research*, *38*(2), 143–162.

Massey, A., & Hill, A. J. (2012). Dieting and food craving. A descriptive, quasi-prospective study. *Appetite*, *58*(3), 781–785.

May, J., & Holton, R. (2012). What in the world is weakness of will? *Philosophical Studies*, *157*(3), 341–360.

Raab, M., de Oliveira, R. F., & Heinen, T. (2009). How do people perceive and generate options? *Progress in Brain Research*, *174*, 49–59.

Robert, P., Onyike, C. U., Leentjens, A. F. G., Dujardin, K., Aalten, P., Starkstein, S., … Byrne, J. (2009). Proposed diagnostic criteria for apathy in Alzheimer's disease and other neuropsychiatric disorders. *European Psychiatry*, *24*(2), 98–104.

Roberts, J. R. (2001). Mental illness, motivation and moral commitment. *Philosophical Quarterly*, *51*(202), 41–59.

Shah, J. Y., & Gardner, W. L. (2008). *Handbook of Motivation Science*. New York: Guilford Press.

Solomon, A. (2014). *The Noonday Demon: An Atlas of Depression*. New York: Simon and Schuster.

Stocker, M. (1979). Desiring the bad: An essay in moral psychology. *Journal of Philosophy*, 738–753.

Widiger, T. A., & Sankis, L. M. (2000). Adult psychopathology: Issues and controversies. *Annual Review of Psychology*, *51*(1), 377–404.