Dual-system theory and the role of consciousness in intentional action

Markus E. Schlosser, University College Dublin, markus.schlosser@ucd.ie

Forthcoming in B. Feltz, M. Missal, & A. Sims (eds.) *Free Will, Causality and Neuroscience*.
This is the author's copy (which may differ from the final print version in minor details).

Abstract

According to the standard view in philosophy, intentionality is the mark of genuine action. In psychology, human cognition and agency are now widely explained in terms of the workings of two distinct systems (or types of processes), and intentionality is not a central notion in this dual-system theory. Further, it is often claimed, in psychology, that most human actions are automatic, rather than consciously controlled. This raises pressing questions. Does the dual-system theory preserve the philosophical account of intentional action? How much of our behavior is intentional according to this view? And what is the role of consciousness? I will propose here a revised account of intentional action within the dual-system framework, and we will see that most of our behavior can qualify as intentional, even if most of it is automatic. An important lesson will be that philosophical accounts of intentional action need to pay more attention to the role of consciousness in action.

1. Introduction

The contemporary philosophy of action has revolved around the notion of intentional action, and it is widely agreed that intentionality distinguishes genuine action from mere behavior and mere happenings. In cognitive and social psychology, human cognition and agency are now widely explained in terms of the workings of two distinct systems (or types of mental processes). System 1, as it is often called, is characterized by processes that are fast, effortless, automatic, and unconscious. Examples include intuitive judgments, the recognition of faces and facial expressions, emotional reactions, fight-or-flight responses, and overlearned routines (such as typing a word, playing an instrument, driving a car, and so on). System 2 is

engaged in processes that are slow, deliberate, controlled, and conscious. Paradigmatic examples are conscious and deliberate reasoning, the solution of novel and difficult problems, and the exercise of self-control. This dual-system framework has been deployed successfully in many areas of empirical research and I will assume, here, that it is on the right track in providing an empirically adequate account of human cognition and agency. (For an extensive overview and review see Evans 2008; see also Sloman 1996, Kahneman 2011, Evans & Stanovich 2013.)

Interestingly, the notion of intentionality does not occupy a theoretical role in the dual-system framework. Occasionally, however, some researchers in this area identify intentional actions with actions that are consciously controlled. Further, in the empirical literature it is often claimed that most of our behavior is driven by System 1 processes that are automatic and not consciously controlled. Given this, we face obvious and pressing questions. Is the philosophical account of intentional action compatible with the dual-system theory? If so, how can the dual-system theory account for intentional action? And what is the role of consciousness? If we simply identify intentional actions with consciously controlled System 2 outputs, we face the unpalatable conclusion that most of our behavior is not intentional, provided that most of our behavior is automatic and driven by System 1, as claimed in the empirical literature. We face, then, the further question of *how much* of our behavior can qualify as intentional within the dual-system framework.

In order to address those questions, we will need to get clearer about the role of consciousness in the dual-system theory and in intentional action. First, I will outline the standard view of intentional action, the dual-system theory, and the role of consciousness in the dual-system-theory. In order to discuss the role of consciousness in intentional action, I will consider five cases in which the activation of a social stereotype influences behavior in different ways and with varying degrees of conscious awareness. This will allow us to see

how intentional action can be captured within the dual-system theory, and it will allow us to see that most of our everyday behavior can qualify as intentional, even if most of it is automatic. In particular, on the basis of the discussion of those five cases, I will a propose revised version of the standard view according to which automatic action (or so called "automatic goal pursuit") can qualify as *derivatively* intentional, if it has an appropriate history of habit formation.[1] An important general lesson will be that philosophical accounts of intentional action need to consider the role of consciousness in action more carefully, both in the guidance of action and in the history of habit formation. Further, the discussion will point to an important distinction between two kinds of intentions, and we will see how intentional action and goal-directed behavior can come apart.

2. The standard view of intentional action

In the philosophy of mind and action, it is generally agreed that intentionality is the mark of genuine agency. For some time, it was also generally agreed that intentional agency can be explained in terms of the roles of the agent's desires and beliefs (largely due to the influence of Davidson 1963). In particular, it was widely assumed that to act intentionally is to act for a reason, and that acting for a reason is to be explained in terms of causation and rationalization by the agent's desires and beliefs. It is still widely held that there is a close connection between intentional action and acting for reasons—that intentional actions are *usually* performed for reasons. But the underlying claim that intentions can be reduced to desires and beliefs is now widely rejected (largely due to the influence of Bratman 1987). According to most contemporary versions of this standard theory of action, intentions play a crucial and irreducible role in practical reasoning, long-term planning, and in the initiation and guidance

[1] For related but nevertheless quite different approaches to this complex of issues see, for instance, Carruthers 2009, Frankish 2009, Evans 2010, and Hommel 2017.

of action. On this view, the intentionality of action is to be explained in terms of the initiation and guidance by *intentions*, construed as irreducible mental states (Bratman 1987, Mele 1992, Enç 2003). In what follows, I will assume this as the current version of the standard view.

Philosophers of action have tended to assume that most of our everyday behavior qualifies as intentional, partly because they have assumed that the standard view can accommodate the intentionality of habitual actions. Davidson, for instance, noted that "we cannot suppose that whenever an agent acts intentionally he goes through a process of deliberation or reasoning" (1978: 85). On his view, actions are intentional only if they are caused and rationalized by the relevant desires and beliefs. But it is not required that the agent consciously considers the relevant desires and beliefs in reasoning. Rather, "if someone acts with an intention, he must have attitudes and beliefs from which, had he been aware of them and had the time, he *could* have reasoned that his action was desirable" (1978: 85). Davidson's example is an agent who adds spice to a stew with the intention of improving the taste. We can certainly imagine that this action is highly habitual, such that the agent's mind is preoccupied with something else. The action would nevertheless seem to be intentional (more on this below, in section 6). Proponents of the view that intentions are irreducible usually offer a similar qualification. According to the current standard view, intentions are defined in terms of their functional roles, which include, most importantly, the initiation and guidance of action (Bratman 1987, Mele 1992, Enç 2003). According to the corresponding qualification, it is not required that the agent is consciously aware of the intention. It is sufficient, that is, if the action is initiated and guided by an intention that is consciously *accessible* (see Mele 2009: Ch. 2).

With this qualification in place, the standard view can accommodate the intentionality of habitual actions. For, even if most of our everyday behavior is habitual, as it seems, it is plausible to assume that most of our behavior is intentional, because it is plausible to assume

that most habitual actions are initiated and guided by intentions that are consciously accessible (such that the agent could tell us his or her intention, if we asked).

3. Dual-system theory

Dual-system theories "abound in cognitive and social psychology" (Evans 2008). They were first proposed to account for biases in logical reasoning, but are now widely deployed in the explanation of human cognition, decision-making, and agency. There are numerous versions of this view, but they share the same basic structure. Until recently, this structure was usually described by means of dichotomies. System 1 processes were characterized as fast, effortless, automatic, and unconscious. System 2 processes were characterized as slow, deliberate, controlled, and conscious. More recently, this way of capturing the core of the dual-system theory has come under attack (for a summary of those criticisms and references see Evans & Stanovich 2013). Proponents of the view have acknowledged some of the criticisms, and they have adjusted the presentation of the view accordingly. In what follows, I will base my discussion on the recent accounts of the theory provided by Evans & Stanovich 2013 and Kahneman 2011.

According to Evans & Stanovich, the mentioned dichotomies capture "correlates" of the two systems, but they should not be taken as definitive (necessary and sufficient conditions). On their current view, the defining features of System 2 processes are "cognitive decoupling" and the "strong loading on the working memory resources that this requires" (2013: 226). This is in line with Kahneman's view, in which System 2 processes are defined in terms of the kind of effort that is required in the simultaneous maintenance of several (and therefore decoupled) representations in working memory, and that is characteristic of conscious problem solving, reasoning, and deliberation (2011: Ch. 2). Evans & Stanovich define System 1 now solely in terms of "autonomy". By this they mean, roughly, that System 1 responses are triggered by their stimulus without the involvement or intervention of System 2 (2013: 236).

This coheres, again, with Kahneman's account, according to which the defining and unifying feature of System 1 processes is their automaticity (2011: Ch. 1).[2]

How do the two systems lead to action? The emerging consensus is that System 2 is endowed with the top-down control to inhibit or override System 1 processes. This is what Evans & Stanovich call a *default-interventionist* architecture.[3] On their account, System 1 usually provides the "default response" to a given task or situation, which arises quickly and intuitively. System 2 may intervene by overriding or inhibiting the default response, if there is sufficient time and if the agent is motivated to engage in effortful System 2 processing. On their view, "most behavior will accord with defaults, and intervention will occur only when difficulty, novelty, and motivation combine to command the resources of working memory" (2013: 237). Kahneman proposes a similar architecture. On his view, System 1 continuously and automatically generates "suggestions" (impressions, desires, and intuitions), which are often endorsed by System 2 "with little or no modification" (2011: 24). System 2 engages in effortful processing and takes over when difficulties arise, when errors are detected, or when System 1 fails to provide a default response. In what follows, I will assume that a default-interventionist architecture provides the correct account of how the two systems lead to action, and the details of this view will be further specified and qualified in due course.

---

[2] Evans & Stanovich (2013) now prefer the terminology of dual-*processes*, because they do not mean to imply that there are two separate systems in the brain. System 1, in particular, is not one unified mechanism, but an array of automatic processes or modules. Kahneman acknowledges this, but he nevertheless keeps the terminology of systems. This is common practice, and I will continue to use the systems terminology as well.

[3] Evans & Stanovich distinguish between two architectures: *parallel-competitive* and *default-interventionist*. According to the former, the two systems simply compete for control, in parallel, and there is no systematic interaction. Evans & Stanovich argue convincingly that this is implausible (2013: 237). On this view, it seems mysterious how System 2 can ever gain control over behavior, because System 1 is always quicker than System 2. Further, System 2 operates with precious working memory resources, and one would expect that such a high-level system is provided with a more systematic access to the motor control system.

4. An obvious but unsuccessful proposal

From what we have seen so far, it seems already clear that the standard view of intentional action and the dual-system theory are not incompatible. According to one obvious proposal on how to locate or capture intentional action within the dual-system framework, intentional actions are simply System 2 outputs: actions that are generated by System 2 processes. It seems that some of the researchers who work on automaticity and dual-system theory hold this view of intentional action. It seems that way, because automatic processes are sometimes *contrasted* with processes that are consciously controlled *and* intentional (see Bargh 1994, Bargh & Chartrand 1999, Bargh 2005, Evans & Stanovich 2013, for instance).[4] As mentioned, the reduction of intentions to desires and beliefs is now widely rejected in philosophy, but it is still commonly assumed that intentions are usually based on desires. One could capture this in accord with the view just outlined by adding the assumption that System 1 generates desires that may or may not be endorsed by System 2, assuming, further, that the endorsement of a desire may consist simply in the formation of an intention to pursue the desired goal.[5]

However, as already indicated, this view has a serious drawback. In the empirical literature, it is often claimed that the great majority of our behavior is automatic. In an influential review article, Bargh & Chartrand (1999) evoked the "unbearable automaticity of being", referring to Baumeister et al. (1998), who estimated that only about five percent of our behavior is consciously controlled. Similarly, Aarts et al. (2005: 466) confidently assert

[4] According to Bargh 1994, the absence of intentionality is one of the "four horsemen" of automaticity. In this relatively early review of automaticity research, Bargh characterized intentionality in terms of conscious control. In later work, Bargh mentions intentionality only in passing, usually as synonymous with conscious control.

[5] The endorsement of a desire may of course involve more than that, such as the reflective judgment that the desired end is desirable or good. The proposed claim here is that endorsement *may* consist only in the formation of the corresponding intention.

that “most of our social behavior occurs in an automatic fashion and originates in the unconscious”. If we identified intentional actions with System 2 outputs, and if we contrasted thereby automatic System 1 outputs with intentional System 2 outputs, we would have to conclude that the great majority of our actions are not intentional. Moreover, the implicit identification of intentional actions with consciously controlled actions is at odds with the mentioned qualification of the standard view, according to which intentional action does not require conscious awareness of the relevant mental attitudes (see section 2). And this shows, in turn, that one could not accommodate the intentionality of habitual actions if one simply identified intentional actions with System 2 outputs.

Further, it is questionable that all desires are generated by System 1 and that the endorsement of desires is always due to System 2. Philosophers have long noted that many of our desires appear to be reason-responsive, and there is now a considerable amount of empirical evidence in support of this view (see Dill & Holton 2014). Some desires simply vanish when one learns that the desired object is unattainable, or when one judges that something else would be better, for instance. Some desires, that is, are responsive to the judgments and intentions generated by System 2, and some desires may even be generated by System 2. Given this, it would simply be a mistake to assume that *all* desires are first generated by System 1, and then, perhaps, endorsed by System 2.

Concerning the endorsement of desires, there is no good reason to think that this is always a System 2 process. System 2 processes are effortful in the sense that they involve cognitive decoupling and a strong load on working memory. It seems clear that we often endorse desires, intuitions, and other “default suggestions” consciously and swiftly, without engaging in effortful reasoning or cognition. Given this, the endorsement of desires need not be a System 2 process. Given, further, that the endorsement of a desire may consist simply in

the formation of an intention to pursue the desired goal, this means that the formation of an intention need not be a System 2 process.

To see why this is plausible, note that many System 1 processes can be flexible, in the sense that their execution can be sensitive to the features of the particular situation and to the agent's beliefs about how to pursue or realize the goal. Examples from priming studies will help to illustrate this point. It has been shown, for instance, that when subjects are primed, in a first task, with words that are related to rudeness, kindness, helping, or so, they are more likely to engage in rude, polite, or helping behaviors in a subsequent task (Aarts et al. 2005, Bargh & Williams 2006).[6] Rude, polite, or helping behaviors are typically sensitive to the features of the particular situation and to the agent's background beliefs about how to pursue the action in the given situation. This kind of flexibility cannot be explained in terms of rigid stimulus-response associations, which map a particular type of input (or situation) to a particular type of output (or action). For this reason, most researchers in this field agree that such actions are instances of "automatic goal pursuit": actions that are to be explained in terms of the activation of goal representations that initiate and guide the execution of the behavior (Custers & Aarts 2010).[7] On this view, the activation of the stimulus leads to the

[6] Recently, some of the seminal experiments in this area have been called into question. Several attempts at replication have either failed or produced only significantly smaller effects. There is, however, also a very large body of research that corroborates the findings. Even if the effects are small and difficult to reproduce, it seems rather unlikely that there are no real effects that underlie the results in this area of research. For a recent discussion of those issues see Open Science Collaboration 2015.

[7] According to Levy (2014: 118, note 5), the evidence does not indicate automatic goal pursuit, because the actions in question are merely *modulated* by the priming. This is a plausible interpretation of many of the classic findings. For instance, in one such experiment, priming influenced the accuracy of the task performance by influencing the speed of the task completion. But it can be argued that even in cases of this kind a non-conscious goal was "superimposed on an already activated parallel conscious task goal" (Bargh et al. 2001: 1024). In other experiments, an interpretation in terms of modulation is rather implausible. For instance, the subliminal priming of cooperation or helping behavior is not plausibly interpreted as a mere modulation (see Bargh et al. 2001, Aarts et al. 2005).

performance of the action indirectly, by way of the activation of the relevant goal, such that the resulting goal-pursuit is sensitive to the agent's beliefs about how that goal is to be pursued or realized in the circumstances. (Direct empirical evidence for this kind of flexibility is provided by Hassin et al. 2009.)

Now, according to one standard definition in psychology, goals are "internal representations of desired states" (Austin & Vancouver 1996: 338). According to another, a goal is the "representation of a future object that the organism is committed to approach or avoid" (Elliot & Fryer 2008: 244). The former corresponds, very roughly, to the philosophical notion of desire. The latter corresponds to the philosophical notion of intention, again very roughly.[8] We might say here that the definition of goals in psychology is ambiguous between the philosophical notions of desire and intention. Or, more positively, we might just as well say that the definition encompasses both the philosophical notion of desire and of intention. However, talk of automatic goal-*pursuit* suggests that the scientists in this particular research area have in mind the formation of a representational state that initiates and guides the pursuit of the relevant action—a state, in other words, that commits the agent to the pursuit of a particular action. Given this, we can say that the scientists in this area agree that automatic goal-pursuit, which is a paradigmatic System 1 process, is to be explained in terms of the automatic activation of *intentions*.

One might, of course, reject a definition of intentions solely in terms of the functional roles of initiation and guidance, and insist, for instance, that genuine intentions must be consciously formed or accessed. But this would not facilitate real progress. For if one simply claimed that intentional action requires conscious intention, without further qualification, one

[8] See Bratman 1987 and Mele 2009, for instance, who stress this element of commitment, or of having settled the question of which action to pursue, which distinguishes intentions from desires.

could not accommodate the apparent intentionality of habitual actions, and one would once again face the undesirable conclusion that most of our actions are not intentional.

We have seen that the standard view of intentional action is clearly not incompatible with the dual-system theory. But a satisfactory account of intentional action within this theory has to be more nuanced than the view considered in this section. In order to make constructive progress on this, we need to get clearer about the role of consciousness in the dual-system theory and in intentional action.

5. The role of consciousness in the dual-system theory

In the previous section, we have seen that consciousness is not a distinguishing feature of System 2, as System 1 processes may also involve consciousness. The default suggestions of System 1 appear often in consciousness, and if one endorses such suggestions by consciously forming a judgment or an intention without effortful reasoning, then the endorsement is itself part of the System 1 process. What is distinctive of System 2 processes is, more specifically, that one consciously conducts individual steps of reasoning in an effortful process that requires the simultaneous maintenance of decoupled representations in working memory. What kind of consciousness is at issue here?

In philosophy, it is common to distinguish between *phenomenal* and *access* consciousness. Phenomenal consciousness is usually characterized in terms of *what it is like* to have a certain experience (such as the distinctive quality of a certain visual or auditory perception). According to common characterizations of access consciousness, a representation is conscious if it is available for reasoning, decision-making, and the control of action (including verbal report). Access consciousness is a functional notion—it specifies the functional roles of conscious mental states.

In a review article, Evans (2008) points out that there is an operational definition of consciousness that is shared, at least implicitly, by most dual-system theories: “System 2

thinking requires access to a central working memory system of limited capacity" and "what we are aware of at any given time is represented in this working memory, through which conscious thinking flows in a sequential manner" (2008: 259). This is a definition in terms of access and it specifies how representations become accessible for further processing—namely, by virtue of being in the central working memory system.

Evans notes, further, that this view is closely associated with the global workspace theory of consciousness. According to this view, a representation is conscious if it is "broadcast" in the "global workspace", which makes it available to a wide range of consumer systems (Baars 1997). Initially, this talk of broadcasting in the global workspace was merely a heuristic metaphor. But more recently this view has been developed into a global *neuronal* workspace theory of consciousness, which specifies concrete neural mechanisms for the workspace and for broadcasting. This theory is empirically well-motivated and it has been successfully deployed in the neuroscientific study of consciousness (Dehaene & Naccache 2001, Baars 2002). There has been some debate about whether or not the workspace can simply be identified with the central working memory system, but it is agreed that working memory is a central component of the workspace (Baars & Franklin 2003, Levy 2014). And as Evans notes, this "association of conscious thought with such a working memory explains the slow, sequential, and low-capacity nature of System 2" (2008: 259).

So, the role of consciousness in the dual-system theory is captured by an operational definition of consciousness that coheres with the philosophical notion of access consciousness and with the global (neuronal) workspace theory of consciousness. But can we plausibly restrict our considerations to access consciousness?

Note, first of all, that we are interested in the role of consciousness in the initiation and guidance of action. That is just to say that we are interested in the functional role of consciousness, which is precisely what a definition of access consciousness is supposed to

capture. Note, moreover, that we are concerned here primarily with the roles of desires, beliefs, and intentions. To insist that phenomenal consciousness must play a role would commit one to the rather implausible view that conscious access to desires, beliefs, and intentions is always accompanied by phenomenal consciousness.

However, one may grant that access consciousness is all that matters in most cases and hold that there are other cases in which mental states influence or motivate behavior in virtue of their phenomenal quality. Suppose, for instance, that certain aesthetic or moral judgments motivate actions in virtue of the fact that they are based on aesthetic perceptions or moral sentiments with certain phenomenal qualities. Would this not show that phenomenal consciousness can play an important and irreducible role in action?

What is at issue in such cases is always the "functional correlate" of the phenomenal experience—that is, the functional role that correlates with the phenomenal quality in question (see Chalmers 1995). It may well be that a mental state has a particular functional role in virtue of having a certain phenomenal quality. But its role in the initiation and guidance of action is still its functional role. Note that there are two different explanatory relations in play here. A mental state plays a role in the initiation and guidance of action *in virtue of* being a mental state with that functional role, and it may have and play this functional role *in virtue of* having a certain phenomenal quality. But its role in action remains its functional role. The explanatory relation between the phenomenal quality and the action is *indirect*, mediated by its functional role, which is the functional correlate of the phenomenal quality in question.

Note, finally, that the functional role of a conscious intention is by no means exhausted by the initiation and guidance of action (including verbal report). In particular, by virtue of being a *conscious* intention it has the functional role of being available for further deliberation

and decision-making, and it plays this functional role by virtue of being broadcast in the global workspace (or central working memory system).

6. The role of consciousness in intentional action

Our main question is how to account for intentional action within the dual-system theory, and we have seen that we need to get clearer about the role of consciousness. As mentioned, in section 2, it is common to qualify the standard view with the claim that intentional action does not *require* conscious deliberation or conscious awareness of the relevant intention. The fact that this is usually *added* as a qualification suggests that it is usually taken for granted that paradigmatic instances of intentional action *are* initiated and guided by conscious intentions (and, perhaps, reasons). I share this assumption, at least as a starting point for the present investigation.[9] In the previous section, we have seen that the role of consciousness in the dual-system theory is captured by the global (neuronal) workspace theory of consciousness, and I have argued that the implicit restriction to access consciousness is unproblematic. With this framework in place, we can now address our main question in a more systematic and nuanced manner. We will consider a series of five cases, in which behavior is influenced by the activation of a social stereotype in five different ways, with varying degrees of consciousness. This distinction between five cases is not exhaustive. But it will cover the full range from automatic and unconscious goal pursuit to conscious and deliberate action, and it will thereby allow us to consider the full range from full System 1 to full System 2 engagement.

As mentioned, experiments show that behavior can be influenced by priming with words that are associated with concepts such as rudeness, politeness, or helping. Likewise, it

[9] This assumption is in line with commonsense or folk intuitions about intentional action. The empirical evidence provided by Malle & Knobe 1996 and Malle et al. 2000 suggests that, according to the folk concept, an action is intentional and based on reasons only if the agent is aware of the relevant intention and reasons.

has been shown that behavior can be influenced by priming with words that are associated with social stereotypes concerning race, gender, or social class, and the same has been shown for the activation of such stereotypes by the presence of group members and perception of group features, such as skin color (Bargh & Chartrand 1999, Aarts et al. 2005, Bargh & Williams 2006). Generally speaking, stereotyping involves generalization by way of categorization and association, and social stereotypes are commonly defined as "generalizations about the shared attributes of a group of people" (Judd & Park 1993). For instance, a person is perceived as Asian, female, or working class, due to certain superficial features, and this categorization is associated with features such as being good at math, bad at driving, or the like.[10] A stereotype is said to be *activated* when the perception or activation of one feature automatically activates an associated feature (or features). Experiments have shown that stereotypes can be activated without the subject's awareness, and they have shown that both the conscious and the unconscious activation of stereotypes tends to influence behavior. In the empirical literature, it is generally assumed that the influence on behavior may be mediated by one of the following two mechanisms (Bargh et al. 2001, Aarts et al. 2005, Bargh 2005, Bargh & Williams 2006). First, stereotypes may become associated with behavioral tendencies, such that the activation of a stereotypical feature automatically activates directly the associated behavioral tendency. Second, it is now assumed that stereotypes may also become associated with goal representations, such that the activation of the stereotype leads to behavior by activating the goal and subsequent goal pursuit. The main difference between the two mechanisms concerns behavioral flexibility. The first mechanism is often described as rigid, in the sense that the activation of the stimulus directly activates a tendency to perform a certain type of behavior. The second mechanism is flexible, in the

[10] Social stereotypes need not be negative, and they need not be inaccurate. But they tend to be oversimplified overgeneralizations (Judd & Park 1993).

sense that it is sensitive to background beliefs about how to pursue or realize a goal in a given situation. To use the example from above, the activation of a social stereotype may become associated with the tendency to perform a certain type of rude behavior. Or, according to the second mechanism, the activation of the stereotype may automatically activate the goal to be rude, such that the pursuit of that goal is guided by background beliefs about how the goal is to be pursued in the given situation. The empirical evidence suggests that both mechanisms can operate entirely without conscious awareness. This evidence on stereotype activation is an important part of the evidence on implicit bias (Greenwald & Banaji 1995, Petty et al. 2008, for instance). The discussion of such cases will make it clear that an account of intentional action within the dual-system theory provides also a useful and plausible framework for how to interpret and diagnose the pernicious influence that implicit biases can have on our behavior.

Before we proceed, let me stress that the main purpose of considering the following five cases is to explore the various possibilities in the theoretical landscape, as it were—possibilities concerning the role of consciousness in intentional action and concerning the underlying mechanisms. In particular, the distinction between cases 4 and 5 will be based on the distinction between the two *possible* mechanisms of automatic goal pursuit just mentioned. I will provide references to empirical evidence, where possible, but it should be noted that all questions concerning the exact mechanisms and concerning whether, or to what extent, human agency instantiates any of those five cases are *empirical* questions—and most of them will remain *open* empirical questions for some time to come.

Case 1: Full awareness and deliberate action

In this first case, the agent is aware of the stereotype activation and its influence leads to action by way of conscious deliberation. Suppose, for instance, that the agent encounters a member of a stereotyped group and that this automatically activates the associated stereotype.

The activation of the stereotype is broadcast, in the global workspace (or working memory), and this instigates conscious deliberation about how to respond, in this situation. The deliberation is conducted by System 2 and the individual steps of the deliberative process are broadcast. This, we may assume, results in the conscious formation of an intention, which is then executed with conscious awareness.

In this case, the action is initiated and guided by a conscious intention, and this intention is based on conscious deliberation. Everyone would agree, I take it, that the action in this case is clearly intentional in the fullest sense.[11] Note, though, that not all of the mentioned features are necessary to support the verdict that the action is intentional. According to the standard qualification of the standard view (see section 2), initiation and guidance by an accessible intention is sufficient for intentional action. When we turn to case 3 we will see that this is in need of further qualification. But it is uncontroversial that initiation and guidance by a *conscious* intention is sufficient, and that conscious *deliberation* is not required. This is why it would be appropriate to describe the action in this case as intentional *and* deliberate or as intentional *in the fullest sense*.

Note that we do not assume that System 2 *intervenes* by inhibiting a default System 1 response. We may assume that System 2 takes over either because the agent is motivated to engage in deliberation, or because the stereotype activation fails to generate a default response. In either case, the core capacity of System 2 to conduct individual steps of reasoning in working memory is fully engaged, but the process is nevertheless not a *pure* System 2 process. In conscious deliberation, we consider reasons and we evaluate them in accord with certain rules (principles or normative standards). Typically, we are unaware of

[11] This verdict is supported by the standard view of intentional action, as outlined in section 2, and by all its main rivals in the philosophy of action. Further, it is supported by the empirical evidence on the folk concept of intentional action provided by Malle & Knobe 1996 and Malle et al. 2000.

why or how we retrieve certain consideration as reasons and we are unaware of why or how we select the underlying rules. The relevant reasons simply appear in consciousness, and the relevant rules are usually operative in the background. The retrieval of reasons and the selection of rules has to be conducted by System 1, even in fully conscious deliberation. Otherwise, we would face a regress of consciously choosing reasons and rules, and consciously choosing reasons and rules for choosing *those* reasons and rules, and so on. Given this, it is clear that no cognitive process can be a pure System 2 process.

Case 2: Full awareness and no deliberation

Suppose now that the agent is aware of the stereotype activation, but that System 2 does not engage in conscious deliberation. As before, the stereotype is activated and its activation is broadcast (in the global workspace or working memory). But this time, the stereotype activation is not followed by conscious deliberation. Rather, it directly activates an associated goal representation (such as the goal to be rude, polite, or to engage in helping behavior). We may assume that the activation of the goal is also conscious, in the sense that it is broadcast, and we may assume that the goal is endorsed by the conscious formation of an intention, which is then executed with conscious awareness.

The action is initiated and guided by a conscious intention. This intention is not based on conscious reasoning, although it may be based on reasons (which are accessible but not accessed in the situation). Either way, the action is clearly intentional, because it is initiated and guided by a conscious intention.[12]

This case does not involve genuine System 2 processing. The agent is aware of the stereotype activation and the goal activation, and the goal is endorsed by the conscious

[12] As in case 1, this verdict is supported by the standard view, its main rivals in the philosophy of action, and the empirical evidence on the folk concept of intentional action provided by Malle & Knobe 1996 and Malle et al. 2000.

formation of an intention. But an engagement of System 2 would require, in addition, that individual steps of effortful reasoning are carried out in the central working memory system.

The action in this case is clearly intentional. I would suggest that our confidence in this verdict is to be explained, in part, by the assumption that the action is initiated and guided by a *conscious* intention. This will become clearer when we turn to the next case, where we will see that matters are far from straightforward once we remove the assumption that the action is initiated and guided by a conscious intention.

Case 3: Goal pursuit and partial automaticity

As in case 2, the stereotype is activated and broadcast, and this automatically activates an associated goal. But suppose now that the goal activation is not broadcast. The agent, that is, is not aware of the goal activation and is therefore not in a position to endorse (or inhibit) the goal pursuit by forming a conscious intention. The empirical evidence suggests that such automatic goal activation may nevertheless result in automatic goal pursuit (for reviews see Bargh & Chartrand 1999, Bargh & Williams 2006, Custers & Aarts 2010). Let us suppose, then, that the goal activation initiates and guides an action without conscious awareness.

How should we judge this case? Is the action intentional? In this case, we need to know more. In particular, we need to know whether or not the performance of the action is habitual. To see this, return again to Davidson's example. The agent adds spice to the stew, mindlessly and without conscious awareness (as we are implicitly invited to assume). Why does it nevertheless seem that the action is intentional? The reason, I suggest, is twofold. First, we assume, with Davidson, that the relevant mental attitudes are accessible—if asked, the agent could readily declare his or her intention (to add spice in order to improve the taste). Second, we are clearly led to believe that the action is habitual and that the agent has the right history of habit formation. Clearly, the agent is not adding spice for the first time, and also not just for the second or third time. We would assume, rather, that the agent has done this over and

over again. And we would assume that the agent did so on some occasions in the past with conscious intent.

With those background assumptions in place, we judge, readily, that the action is intentional. In particular, we implicitly assume that there is an underlying habit and an appropriate history of habit formation. There are, I suggest, two main types of such an *appropriate* history. Either the habit has been *formed* by several performances of the action with conscious intent. Or, it may be that the agent has acquired the habit in some other way, such as by imitation, but has later *endorsed* the action and its goal by forming a conscious intention. In either case, the intentionality of later manifestations of the habit is *derivative*: the intentionality of the action derives from earlier instances of acting with conscious intent.

Those considerations apply, *mutatis mutandis*, to our present case 3. If there is an appropriate history of habit formation, such that the habit was either formed or endorsed by acting on the stereotype activation with conscious intent, then a later manifestation, as in case 3, is derivatively intentional—provided, as we should add, that the intention is still consciously accessible.

The assessment of such cases becomes more difficult once we remove the assumption that there is an appropriate history of habit formation. Davidson's case would simply appear to be very odd without this assumption. Why on earth would one add spice to a stew automatically and without awareness if one had no corresponding history of habit formation? In case 3, in contrast, there is a possible explanation of how and why the agent acts automatically and without awareness. In the literature on implicit bias, it is often assumed that we may acquire the relevant stereotype-goal associations without a history of habit formation (or conscious endorsement), because it is assumed that we may acquire such associations by being exposed to them in our socio-cultural environment. This would explain the association between the stereotype and the goal in our present case 3.

So, the stereotype is activated, and the agent is aware of this. But the agent is unaware of the goal activation and of the subsequent initiation and guidance of behavior. Is the action intentional? I am inclined to think that the action is not intentional.[13] But there is one complication that needs to be addressed here. We assume that the stereotype activation influences behavior by automatically activating a goal representation. A goal representation is, by definition, a mental state that can initiate and guide action. It has, that is, the functional role of an *intention* in the initiation and guidance of action. And this may be taken to suggest that the action is intentional. But what this shows, really, is that we need to be more careful, and that we need to make a distinction. We distinguished already between the automatic activation of a goal representation and the endorsement of such a goal by the conscious formation of an intention. This distinction is by no means *ad hoc*. One can find various incarnations of it in the empirical literature, and evidence from brain imaging studies suggests that the workings of automatic goal activations and conscious intentions are implemented by distinct regions in the brain (Frith et al. 2000, Bargh 2005, Pacherie & Haggard 2010, for instance). There is, that is, good reason to think that the distinction is real, but the terminology is optional. One may, for instance, reserve the term "intention" for mental states that have been consciously formed. Or one may introduce a technical term in order to distinguish between two *kinds* of intentions. For instance, in the empirical literature it is common to distinguish between conscious intentions and "motor intentions". That is nothing other than the distinction between conscious intentions and goal representations (sometimes also called "motor representations").

[13] As before, this intuition is in line with the standard view and with empirical evidence on the folk concept of intentions action (see footnote 11). Note that if an agent has acquired the stereotype-goal association by exposure, then the agent is probably unaware of the underlying association. That is, the agent is not only unaware of the goal activation in the particular case. But the agent has probably never been aware of the fact that the activation of the stereotype activates an associated goal.

So, goal representations may be classified as forming a kind of intention. But, of course, this terminological decision should not lead one to conclude that all actions that are initiated and guided by goal representations are therefore intentional. What matters is not the terminology, but the two substantive issues that have already been discussed: Is the action habitual? Is the agent aware of the goal activation? Given this, we can uphold the proposed claim that the action in case 3 is not intentional, if it is not habitual, and if the agent is not aware of the goal activation and its influence on subsequent behavior.

Note, the suggestion is *not* that the performance of an intentional action requires initiation and guidance by a *conscious* intention. Rather, the suggestion is that if an action is initiated and guided by an automatically activated and unconscious goal representation, and if the action has no appropriate history of habit formation (or conscious endorsement), then the action is not intentional. In such cases, the action is goal-directed, but not intentional.

Note, finally, that System 2 is not involved in the *performance* of the action in any of the variations of case 3. But if the action has a history of habit formation (or conscious endorsement), then System 2 may have been involved on past occasions, depending on whether or not those occasions involved effortful deliberation.

Case 4: Goal pursuit and full automaticity

Assume now that the stereotype and the associated goal are activated automatically, and that neither the stereotype nor the goal is broadcast. As before, the agent is not in a position to endorse (or inhibit) the goal by forming a conscious intention. The empirical evidence suggests that the automatic activation of the goal may initiate and guide goal pursuit without conscious awareness (Bargh & Chartrand 1999, Aarts et al. 2005, Bargh & Williams 2006, Custers & Aarts 2010).

This is an example of fully automatic and unconscious goal pursuit, and the agent is unaware of the stereotype activation as well. As in case 3, we need to consider the agent's

history in order to arrive at a judgment concerning intentionality. It seems, again, that if there is an appropriate history of habit formation or conscious endorsement, then the action is derivatively intentional, despite the fact that the agent executes the action in the present case without awareness. In support of this, consider the following. Suppose that when you drive to work in the morning, you take a right turn at the first junction. Sometimes when you approach the junction, you are aware of approaching a junction at which you have to turn right, and you then automatically prepare for taking the turn. On other occasions, you do not become aware of approaching a junction at which you have to turn right, but you automatically prepare for taking the turn all the same. In the first case, you are aware of the stimulus, in the second you are not. This is analogous to the difference between cases 3 and 4. In case 3 the agent is aware of the stereotype activation, in case 4 the agent is unaware. The analogy suggests that this difference does not make a difference concerning intentionality. In each of those four cases, the action is derivatively intentional if it is the manifestation of a habit with the right history (provided that the relevant intention is still accessible). And if there is no such history, the action does not seem to be intentional. Further, as in case 3, the action is goal-directed, but not intentional. System 2 is not involved in the performance of the action, but it may have been involved on past occasions, in the acquisition or endorsement of the relevant habit.

Case 5: Full automaticity and no goal activation

As mentioned, the activation of a stereotype may influence behavior either by activating a goal or by activating more directly an associated behavioral tendency. Let us consider, then, a final case in which the automatic and unconscious activation of a stereotype influences behavior by way of automatically activating an associated behavioral tendency, and suppose that the agent is unaware of all this (for empirical evidence see Bargh & Chartrand 1999, Aarts et al. 2005, Bargh & Williams 2006, Custers & Aarts 2010).

As in case 4, the agent is unaware of both the stereotype activation and its influence on behavior. And as in cases 3 and 4, we need to consider whether or not the action is the manifestation of a habit with the right history. And as the difference in the underlying mechanisms appears to be irrelevant, our assessment of intentionality should follow our discussion of cases 3 and 4. If the action is the manifestation of a habit with the right history, then it is derivatively intentional. If it is not habitual, then it is not intentional, because the agent is entirely unaware of being influenced by the activation of a stereotype. As in cases 3 and 4, System 2 is not involved in the performance of the action, but it may have been involved on past occasions, in the acquisition or endorsement of the relevant habit.

This case shows, though, that there is a sense in which an intentional action may not be goal-directed—namely, in the sense that it is not initiated and guided by a goal representation. But there is, nevertheless, derivative goal-directedness in such cases, because such actions must be derivatively intentional, if they are intentional at all.

7. The standard view revised

The discussion of those cases made it clear, I think, that the standard view of intentional action is in need of modification and further qualification. Consider, then, the following revised version of the standard view, which summarizes the conclusions and suggestions from the previous section (in a rough and ready fashion).

An action is intentional if and only if:

Either (1) the performance of the action is initiated and guided by a conscious intention.

Or (2) the action is derivatively intentional.

An action is derivatively intentional if and only if:

Either (2a) the action is the manifestation of a habit that has been formed by performing actions of this type with conscious intent (and this intention is still consciously accessible and would still be endorsed).

Or (2b) the action is the manifestation of a habit that has been acquired in some other way and has later been endorsed by the conscious formation of an intention (and this intention is still consciously accessible and would still be endorsed).

According to the standard qualification of the standard view (see section 2), intentional action does not require that the relevant mental attitudes are consciously accessed—accessibility is sufficient. According to the proposed revision, conscious access is also not necessary, but accessibility is not sufficient. The agent must either have a conscious intention that initiates and guides the action (to satisfy 1), or the agent must have consciously formed the relevant intention at some point in the past (to satisfy 2a or 2b). So, according to this account, intentional action does depend on conscious intention, but the performance of particular intentional actions does not. The account makes no explicit mention of Systems 1 and 2. But the discussion of the five cases in the previous section made it clear why and how this account of intentional action can be captured within the dual-system theory (in conjunction with the global workspace theory of consciousness).

It is important to note that this account does not require that the agent has at any point the conscious intention to acquire the habit. The condition on habit formation (2a) requires that the habit is formed by performing the *action* with conscious intent. This requires, typically, that the action is performed or practiced repeatedly, but not that the agent has the conscious intention *to acquire the habit*. Note, further, that the account does not entail that the endorsement of a habit (in 2b) must itself be a mental action. The condition requires that the habit is endorsed by the formation of a conscious intention, and the formation of a conscious intention need not be an action. Arguably, the formation of an intention is an action only if it

is motivated by a further desire or intention (such as the desire or intention to settle the practical question at hand). It seems perfectly possible that the formation of some intentions is not motivated in this way, and it seems perfectly possible that we may consciously acquire intentions in this passive or "non-actional" manner (for more on this see Mele 2003: Ch. 9).

8. Lessons and conclusions

There is no straightforward way to locate intentional action in the dual-system theory, such as by identifying intentional actions with System 2 outputs. But we have seen that the standard view of intentional action can be captured within the dual-system theory, and it has emerged that doing so suggests plausible modifications and qualifications of the view. Further, we have seen that it is important to distinguish between goal representations and conscious intentions, or, alternatively, between two kinds of intentions.

An important lesson is that philosophical accounts of intentional action need to pay more attention to the role of consciousness in action. I have suggested that intentional action depends on consciousness. Consciousness, that is, must play a role at some point, either in the initiation and guidance of the action or during the formation or endorsement of the relevant habit. This does not mean that intentional action depends on the involvement of System 2, because the conscious endorsement of a goal need not involve System 2 processing, as I have argued.

What can we say about *how much* of our behavior can qualify as intentional? We can conclude that most of our everyday behavior may well be intentional, even if most of it is automatic, because most automatic actions may well be habitual and derivatively intentional. This, I should note, includes another way in which automatic actions can qualify as derivatively intentional. Many automatic actions are sub-routines that are in the service of consciously accessible goals and intentions. Common examples are the sub-actions that one performs while playing an instrument or while driving a car. This kind of derivative

intentionality is included, because the initiation and guidance of such sub-routines is always habitual.

To conclude, then, we have seen that the philosophical standard view of intentional action can be captured within the dual-system theory, and we have seen that doing so offers important lessons on how to think about the role of consciousness in intentional action. And we have seen that the findings from the empirical research on automaticity are not so "unbearable" after all, because they do not undermine the assumption that most of our everyday behavior can qualify as intentional.

Acknowledgements

A draft of this paper was presented at workshops and conferences at University College Dublin, University College Cork, UC Leuven, and the HU Berlin. Many thanks for the questions and comments that have helped to improve the manuscript, and special thanks to Francesca Bunkenborg for the very helpful commentary presented at the HU Berlin.

References

Aarts, H., Chartrand, T.L., Custers, R., Danner, U., Dik, G., Jefferis, V.E., & Cheng, C.M. (2005). Social stereotypes and automatic goal pursuit. *Social Cognition* 23: 465–490.

Austin, J.J., & Vancouver, J.B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, 120: 338–375.

Baars, B.J. (1997). *In the Theatre of Consciousness*. Oxford: Oxford University Press.

— (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences* 6: 47–52.

Baars, B.J. & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences* 7: 166–172.

Bargh, J.A. (2005). Bypassing the will: Toward demystifying the nonconscious control of social behavior. In R.R. Hassin, J.S. Uleman & J.A. Bargh (Eds.). *The New Unconscious*. Oxford University Press, pp. 37–58.

Bargh, J.A., & Chartrand, T.J. (1999). The unbearable automaticity of being. *American Psychologist* 54: 462–479.

Bargh, J.A., Gollwitzer, P.M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology* 81: 1014–1027.

Bargh, J.A., & Williams, E.L. (2006). The automaticity of social life. *Current Directions in Psychological Science* 15: 1–4.

Baumeister, R.F., Bratslavsky, E., Muraven, M., & Tice, D.M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* 74: 1252–1265.

Bratman, M.E. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Carruthers, P. (2009). An architecture for dual reasoning. In J.St.B.T. Evans & K. Frankish (Eds.). *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press, pp. 109–127.

Chalmers, D.J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2: 200-19.

Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science* 329: 47–50.

Dancy, J., (2000). *Practical Reality*. Oxford: Oxford University Press.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, 60: 685–700; reprinted in D. Davidson (1980). *Essays on Actions and Events*. Oxford: Clarendon Press, pp. 3–20.

— (1978). Intending. In Y. Yovel (Ed.), *Philosophy of History and Action*, Dordrecht: D. Reidel; reprinted in D. Davidson (1980). *Essays on Actions and Events*. Oxford: Clarendon Press, pp. 83–102.

Dehaene S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79: 1–37.

Dill, B., & Holton, R. (2014). The addict in us all. *Frontiers in Psychiatry* 5: 1–20.

Elliot, A. J., & Fryer, J. W. (2008). The goal concept in psychology. In J. Shah & W. Gardner (Eds.), *Handbook of Motivational Science*. New York: Guilford Press, pp. 235–250.

Enç, B. (2003). *How We Act: Causes, Reasons, and Intentions*. Oxford: Oxford University Press.

Evans, J.St.B.T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59: 255–278.

Evans, J.St.B.T. (2010). *Thinking Twice: Two Minds in One Brain*. New York: Oxford University Press.

Evans, J.St.B.T., & Stanovich, K.E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science* 8: 223–241.

Frankish, K. (2009). Systems and levels: Dual-system theories and the personal-subpersonal distinction. In J.St.B.T. Evans & K. Frankish (Eds.). *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press, pp. 89–107.

Frith, C.D., S. Blakemore, & Wolpert, D.M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London B* 355: 1771–1788.

Greenwald, A.G., & Banaji, R.M. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102: 4–27.

Hassin, R.R., Bargh, J.A., & Zimerman, S. (2009). Automatic and flexible: The case of non-conscious goal pursuit. *Social Cognition* 27: 20–36.

Hommel, B. (2017). Consciousness and action control. In T. Egner (Ed.). *The Wiley Handbook of Cognitive Control*. Oxford: Wiley-Blackwell, pp. 111–123.

Judd, C.M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review* 100: 109–128.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Levy, N. (2014). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Malle, B.F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology* 33: 101–121.

Malle, B.F., Knobe, J., O'Laughlin, M.J., Pearce, GE., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *Journal of Personality and Social Psychology* 79: 309–326.

Mele, A.R. (2003). *Motivation and Agency*. Oxford University Press.

— (2009). *Effective Intentions: The Power of Conscious Will*. Oxford: Oxford University Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science* 349: acc4716-1–aac4716-8.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition* 107: 179–217.

Pacherie, E. & Haggard, P. (2010). What are intentions? In L. Nadel & W. Sinnott-Armstrong (eds.), *Conscious Will and Responsibility: A Tribute to Benjamin Libet*. Oxford University Press, pp. 70–84.

Petty, R.E., Fazio, R.H., & Brinol, P. (Eds.). (2008). *Attitudes: Insights From the New Implicit Measures.* New York: Psychology Press.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin* 119: 3–22.