**DRAFT: published in *Mental Action and the Conscious Mind*, M. Brent and L.M. Titus (eds.), Routledge, 2022.**

**Disappearing agents, mental action, rational glue[[1]](#footnote-1),[[2]](#footnote-2)**

Joshua Shepherd

**Abstract**

This chapter revolves around the problem of the disappearing agent. Shepherd suggests that as typically formulated, the problem relies on an improper focus upon the causation of action, and an inadequate characterization of agency. One result is that a key function of mental action for human agents tends to be misconstrued. Furthermore, Shepherd argues that an adequate characterization of agency illuminates why agents may seem (misleadingly) to disappear in some cases of action, and illuminates as well a key function of mental action for human (and similar) agents. According to the sketch offered in this chapter, mental action enables human agents to rationally glue themselves together.

**1. How agents disappear**

The problem of the disappearing agent arises, for some, in the context of reflection upon event-causalist accounts of intentional action (Velleman, 1992; Wallace, 1999; Brent, 2017).[[3]](#footnote-3) According to such accounts (Brand, 1984; Mele, 1992), intentional action is behavior caused in the right way by the acquisition of certain mental states (e.g., intentions). The worry is that talk of events and states within an agent cannot amount to action by the agent.

1.1 Velleman

David Velleman complains that the ‘standard story’ of intentional action leaves something out: “my objection is that the occurrences [the standard causalist view] mentions in the agent are no more than occurrences in him, because their involvement in an action does not add up to the agent’s being involved” (1992, p. 463).

The standard causalist has an obvious response, and Velleman considers it. It is that the standard story contains the agent implicitly:

The reasons, intention, and movements mentioned in the story are modifications of the agent, and so their causal relations necessarily pass through him. Complaining that the agent takes no part in causal relations posited between reasons and intention, they might claim, is like complaining that the ocean takes no part in causal relations posited between adjacent waves. (Velleman, 1992, p. 463)

Velleman rejects this response, offering two cases.[[4]](#footnote-4) He first references Frankfurt’s unwilling addict. This is a person who has a first-order desire in favor of taking a substance, and a second-order desire in favor of dropping the first-order desire. The addict takes the substance because of the first-order desire’s strength. The addict takes it unwillingly because the action runs counter to the second-order desire. Velleman comments that “being the subject of causally related attitudes and movements does not amount to participation of the sort appropriate to an agent” (1992, p. 463). Velleman’s view is that an account of action that includes only mental states and causal relationships between them makes the agent disappear – there is a thing called an agent, it participates in action, this participation only takes place when standards are met, and these standards go beyond the normal (non-deviant) causal operations of normal mental states (e.g., intentions and beliefs).

Velleman further illustrates this view with a different kind of case. In this case a character meets with an old friend “for the purpose of resolving some minor difference” (1992, p. 464). Things have been complicated, such that during this meeting subconscious intentions to sever the friendship ‘crystallize,’ and the meeting ends with both characters angry at each other. Velleman intends this as a case in which action occurs – there is non-deviant causation of behavior by appropriate mental states. Even so, in this case the agent disappears. Velleman claims: “Surely, I can believe that the decision, though genuinely motivated by my desires, was thereby induced in me but not formed by me; and I can believe that it was genuinely executed in my behavior but executed, again, without my help” (1992, pp. 464–465).

This case is different in important ways from that of the unwilling addict. What they have in common is that both cases capitalize on *agentive complexity* and the *internal practical conflict* that may result. The addict has desires at cross-purposes. The angry friend has intentions at cross-purposes – the explicit intention to meet in order to resolve a difference, and the subconscious intention to end the friendship.

Velleman’s proposal is sensitive to the need to resolve internal practical conflict.[[5]](#footnote-5) He argues that we need a mental state to play the functional role appropriate to the agent’s participation in action. This role is that of reflecting on available reasons for action, and taking sides with the best reasons. Velleman then argues that the agent is involved in the action when these mental actions of reflection and taking sides are motivated by a special kind of mental state – a desire to act in accordance with reasons.

We say that the agent turns his thoughts to the various motives that give him reason to act; but in fact, the agent’s thoughts are turned in this direction by the desire to act in accordance with reasons. We say that the agent calculates the relative strengths of the reasons before him; but in fact, these calculations are driven by his desire to act in accordance with reasons. We say that the agent throws his weight behind the motives that provide the strongest reasons; but what is thrown behind those motives, in fact, is the additional motivating force of the desire to act in accordance with reasons. For when a desire appears to provide the strongest reason for acting, then the desire to act in accordance with reasons becomes a motive to act on that desire, and the desire’s motivational influence is consequently reinforced. The agent is moved to his action, not only by his original motive for it, but also by his desire to act on that original motive, because of its superior rational force. This latter contribution to the agent’s behaviour is the contribution of an attitude that performs the functions definitive of agency; it is therefore, functionally speaking, the agent’s contribution to the causal order. (Velleman, 1992, p. 479)

Very few have accepted Velleman’s proposal. Later I discuss what I think has gone wrong. But there is something interesting and insightful here. When Velleman reaches for the role of the agent, he reaches for certain kinds of mental actions – deliberation, decision, reflection upon reasons broadly construed (which is plausibly constituted by mental actions of many types – counterfactual thinking, prospective imagination, attempts to remember alternatives, attempts to infer likely consequences, etc.). These are mental actions that have a function of navigating internal practical conflict.

1.2 Wallace

Let us bookmark this point, and turn to R. Jay Wallace’s (1999) discussion of the disappearing agent. Like Velleman, he is concerned with cases of internal practical conflict, and these cases stem from agentive complexity of the sort humans exemplify. For Wallace is motivated by cases of akrasia, and in particular by cases of action motivated by addiction.

Wallace contrasts two models of the will. He calls the first the hydraulic model. This model depicts desires “as vectors of force to which persons are subject, where the force of such desires in turn determines causally the actions the persons perform” (Wallace, 1999, p. 630). The concept of desire in play excludes intentions – one can desire something without intending or choosing it. But on the hydraulic model the agent always does what she most desires to do. In this way desire plays the key role in causal explanations of action.

They determine which action we perform by causing the bodily movements that we make in acting, the assumption being that the strength of a given desire is a matter of its causal force in comparison to the other given desires to which we are subject. (Wallace, 1999, p. 631)

Wallace’s primary complaint about this model is that it “leaves no real room for genuine deliberative agency” (1999, p. 633). Cases of akrasia motivate this complaint. For on the hydraulic model, there is no room to maintain that the agent could have done otherwise than act against her best judgment, unless the strength of her desires were different. Wallace comments: “*Given* the causal strength of the various desires to which they are actually subject, together with their actual beliefs, it turns out that *akratic* agents simply lack the capacity to do what they judge best” (1999, p. 633).

So, Wallace complains that a picture of agency that does not permit agents to rationally handle cases of internal practical conflict – a desire goes one way, while a judgment goes another, and the rational move is to follow the judgment – makes the agent disappear. Citing Velleman, Wallace says the following: “Action is traced to the operation of forces within us, with respect to which we as agents are ultimately passive, and in a picture of this kind real agency seems to drop out of view” (1999, p. 633).

Recall that at this stage Velleman posited a novel kind of desire to stand in for the agent by motivating a range of rationality-promoting mental actions. Wallace’s proposal is slightly harder to parse. It involves a different motivational state, and it appears to involve, as well, a novel capacity. According to Wallace’s *volitionalist model*, the special motivational state is a volition – “a kind of motivating state that … [is] directly under the control of the agent” (1999, p. 636). Wallace offers intentions, decisions, and choices as sub-types of volitions. So, these special motivational states are defined in terms of a special capacity – the agent’s direct control over them. It is in these states and this capacity, according to Wallace, that we find agency itself:

[I]ntentions, decisions, and choices are things we do, primitive examples of the phenomenon of agency itself. It is one thing to find that one wants some chocolate cake very much, or that its odor reminds one of one’s childhood in Detroit, quite another to resolve to a piece. The difference, I would suggest, marks a line of fundamental importance, the line between the passive and the active in psychological lives. (1999, p. 637)

One should worry that appeal to an agent’s direct control over a special class of motivational states obscures more than it illuminates. How are we to understand the capacity for direct control? Is it always rational? If so, then it is redundant – judgments about what it is best to do would seem to be enough. If not – and since it is a capacity humans possess, this is the option we have to take – we lack an explanation for why it operates rationally at times, and irrationally at others.

But detailed criticism of Wallace’s proposal is not my aim here. It is the pattern of reasoning that interests me.

1.3 Brent

Like Velleman and Wallace, Michael Brent looks at event-causalist accounts of intentional action and sees disappearing agents. He is not motivated, however, by cases involving agentive complexity or internal practical conflict. It seems that no special rational state or capacity classified in terms of event causation will satisfy him. Brent argues that we need a new view of causation. According to Brent, it is not states or events that act, it is agents, and no event-causal account could “properly account for your causal role when you are initiating, sustaining, and controlling the movements of your body during an action” (2017, p. 665). What is required to capture this is “a plausible alternative conception of causation” (Brent, 2017, p. 668).

But this does not seem to be all. Brent endorses non-reductive agent causation. But he also endorses the posit of a special causal power that amounts to agent causation. This causal power manifests in the exertion of effort.

I suggest that the fundamental difference between the bodily movements that you are making happen during an action and those movements that are merely happening is that the former are occurring in conjunction with your exertion of effort, whereas the latter are not. Although the bodily movements you are making happen during an action and those that are merely happening might seem to be instances of the same type, they are not. The causal contribution of your exertion of effort differentiates those movements that you are making happen while acting from those that are merely happening, marking the movements as categorically distinct. (Brent, 2017, p. 667)

As with Velleman and Wallace, one might worry that this proposal raises more problems than it solves. We lack reasons to think that effort uniquely produces intentional actions, as opposed to a wide range of effects. Nothing about a notion of substance causation in general could help here, since even if substance causation is fundamental, lots of non-agentive substances cause lots of non-actional effects. And there seems nothing especially illuminating in the notion of effort itself. It seems plausible that effort could sometimes cause non-actional byproducts. (The event-causalist might suggest that we introduce intentions here, to distinguish between effects that amount to intentional actions and effects that do not. Brent would reject this move.) We also have reason to think that a problem of deviant causation afflicts this proposal as much as it does event-causalist accounts. For the conjunction of effort and bodily movements have to be stitched together somehow.

Again, though, it is not detailed criticism that is my aim, but the pattern of reasoning at work.

1.4 The moral of the story so far

In my view Velleman’s, Wallace’s, and Brent’s proposals are responsive to a genuine problem, namely, how to understand the relationship between the agent and the actions the agent produces (or executes). Further, both Velleman’s and Wallace’s proposals are sensitive to a deep truth about the nature of agency. The deep truth is the close association of agency and the application of rational standards to behavior.

Like Velleman, Wallace is motivated by cases of internal practical conflict. Such cases, if handled improperly, seem to make the agent disappear. Similarly to Velleman, Wallace posits a special class of mental state that could bring the agent back into the picture. In addition, he commits to a special mental capacity of direct control over these states. Both theorists are, then, guided by an image of the true core of agency.

Brent does not acknowledge the importance of practical norms in understanding the nature of agency. But he is viewing a version of this image as well. For Brent, the true core of agency involves a causal power to produce effects by exerting effort.

All three theorists are, in different ways, committed to capturing the core of agency in terms of some very special item – a state, or capacity, or mode of causation. And all three execute this commitment at the level of the production of individual actions. They seem to want to shoehorn the agent into the action at the level of action explanation. This leads them to posit esoteric items that could play a role in action explanation.

All three theorists, and the many who have found the problem of the disappearing agent compelling, are misled. An alternative picture of agents is required.

**2. The nature of agency**

The problem of action is, quoting Frankfurt, “to explicate the contrast between what an agent does and what merely happens to him” (1998, p. 69). As I understand it, it is a contrast between kinds of event: between *action* on the one hand, and *mere behavior* on the other. I am focused here on a different – though importantly related – problem. We might call it the problem of agency, and understand the challenge as one to explicate the contrast between what qualifies as an agent and what does not. As I understand it, it is a contrast between kinds of system: between *agents* on the one hand, and *non-agential systems* on the other.

A number of philosophers have found attractive some version of the idea that to qualify as an agent a system should conform to certain normative standards. Often this is put in terms of rationality. So, for example, Donald Davidson claims that “An agent cannot fail to comport most of the time with the basic norms of rationality” (2004, pp. 196–197). And Christian List and Phillip Pettit claim that “The very idea of an agent is associated with some standards of performance or functioning, which we call ‘standards of rationality’. These must be satisfied at some minimal level if a system is to count as an agent at all” (2011, p. 24).

Some argue for a more encompassing notion of agency. It will be useful to look briefly at it.

Consider a very simple system, only capable of moving in one direction along a flat surface. It has some internal structure. It has some causal powers. We might impute a function to this system, based on its survival needs: the system needs to find and fall into small gaps in a surface on which it moves. If it fits into the gap, it wins. (Say it avoids predators, or finds food.) The system does so in the only way it can, by moving blindly in one direction along the wall. There aren’t so many small gaps in the wall, but every once in a long while – make it as unlikely as you like – it comes across one. It wins. This may be enough for the system. Such a system does not trend in the direction of agency.

Now consider a slightly more complex system – the paramecium, a single-celled eukaryote. It has some internal structure. It has some causal powers. It moves through certain liquids by the coordinated beating of its cilia. It can navigate around obstacles or escape certain substances by way of an “avoiding reaction” – a reversal of its direction of movement, and a slight changing of course in some random direction (Kung & Saimi, 1982). This is not a very efficient way of navigating, but the thing is stuck with very short cilia. In any case, it is also capable of reproducing, and its methods appear good enough for evolution to keep it employed – many a paramecium has survived long enough to reproduce.

Some think that unlike our earlier system, the paramecium trends in the direction of agency. Tyler Burge argues that “primitive agency” extends down to the level, at least, of single-celled eukaryotes. Burge points to the orientation behavior of such organisms:

Taxes are directional movements with respect to stimulations in the environment. They require sensory capacities that are directional. Usually determining direction depends on there being two or more locations of sensory receptors on the body of the organism. Directional movement is usually achieved by some mechanism in the animal for simultaneous differentiation of intensities of stimulus registration in different bodily sensors. For example, the animal might turn toward or away from the direction perpendicular to the side of its body that receives the most intense stimulus registration. (2009, p. 258)

Burge judges that coordinated, functioning orientation behavior of simple organisms – e.g., “The paramecium’s swimming through the beating of its cilia, in a coordinated way, and perhaps its initial reversal of direction” (2009, p. 259) – qualify them as agents. As Burge writes, “Such organisms are capable of steering toward or away from a stimulus source, subsequent to internal differentiations between stimulus intensities in different areas of the body” (2009, p. 258). The movement toward a stimulus is caused in a different way than the movement away from a stimulus, and the difference makes sense in light of the system’s own activity – the transitions between states of the system that are differentially sensitive to stimulus source and intensity. That is, the system’s behavior is not only reliably produced, it is coherently produced given the circumstances. And it permits something like success. The system’s behavior is related to imputable goals regarding its needs (for safety, for finding energy sources, or whatever) with respect to its environment. In its typical behavioral circumstances, this orientation behavior reliably leads to successful (enough) approximation of these goals.

Many will disagree with Burge that we find agency at this level. After all, reproduction is no less complicated and important a process for the paramecium than is locomotion. But it is less intuitive to think of a paramecium’s asexual reproduction, by a process of binary fission, as an example of primitive agency. That’s just the mechanics of life. And if so, why not the beating of the cilia, or the avoiding reaction (which, by the way, often occurs spontaneously)?

Whatever we think about the agency of a paramecium, Burge is right to emphasize continuity between this level and others. At this low level we find key ingredients of agency. First, behavioral standards – standards of success – must be imputable to the system. This trends in the direction of the application of rational norms. Second, behavior must be coherent in light of the relevant behavioral standards. This trends in the direction of rational behavior. Third, behavior must be reliable in meeting or approximating these standards – the system must succeed, to some degree. This trends in the direction of control.

Look at systems more internally sophisticated (and usually more causally powerful) than the paramecium. Behavioral standards that apply to the system can still be drawn from the system’s needs or functions. But at this level the system begins to set certain standards for itself. For at this level the system has the capacity to represent objects, and goals for action regarding these objects – to token psychological states that link it and its behavior to the world in reliable ways. Burge here invokes the notion of a perspective: “When perception sets an object for animal action, agency reaches a new level of sophistication. The action is suited to a goal that the animal itself perceptually represents. If an animal can perceive, it has some perspective on its objectives” (2009, p. 267).

One might think that this is the level at which agency truly emerges. This is what Sean Thomas Foran (1997) argues. According to Foran, an animal moves itself, as opposed to being passively moved as a rock is, when the animal’s movements are shaped with respect to objects of that animal’s perception. Foran’s notion of movement being shaped seems similar to the notion I offered just above, of a system’s behavior being coherently produced.

“Movement shaped with respect to an object of perception” does not simply mean “movement caused by perception.” Movement can be caused, in some quite general sense of “caused,” by perception without being shaped with respect to the object of that perception. Consider this example. Suppose that when a certain kind of quadruped animal sees one of its natural predators, it immediately lowers itself to the ground and remains still. Perceiving the predator causes the animal to lower itself, but the movement that is caused is not shaped with respect to the predator. The movement is still shaped with respect to something the animal perceives, the ground, but its perception of the ground is not what led it to lower itself: this episode of movement was caused by perceiving the predator. (1997, pp. 41–42)

At this level, perhaps, it becomes appropriate to think of coherent production of behavior in terms of practical rationality. When a system can represent behavioral targets, and implement plans for behavior that approximate standards of success regarding these targets, that system’s behavior might well be considered practically rational. And some of that system’s behavior might be considered intentional action.

We are still, however, at a level of relative simplicity. At this level it is important that the system be embedded in circumstances in the right ways. For, while the system may be able to represent targets for behavior and deploy plans to hit these targets, the behavioral profiles deployed in following the plan may be *inflexible*. And inflexible behavioral profiles contain a flaw regarding the meeting of certain behavioral standards.

Distinguish between success according to the standard a system’s particular goal or plan sets, and success according to the standards that apply to that system more broadly. If the system is at all complex, then the standards that apply to it will be broader than the standards a particular goal or plan sets. It will have a range of needs, or perform a range of functions. It may even have a range of intentions, which need to be delicately executed in order not to fail with respect to some of them. Inflexible behavioral routines lock the system into one way of behaving, making it difficult for the system to change tack, or to adjust even slightly. As a result, any infelicitous circumstances, or any kinks in the plan, may throw the system off course.

Consider the digger wasp, *Sphex ichneumoneus*. In preparing to lay her eggs, the Sphex displays some extraordinarily intelligent-seeming behavior. It catches and drags a cricket into its burrow, lays eggs in the burrow, closes the burrow, and leaves.

Regarding this behavior, Woolridge (quoted in Dennett, 1984, p. 11) comments (though the actual details regarding Sphex behavior may be more complicated – see Keijzer, 2013):

To the human mind, such an elaborately organized and seemingly purposeful routine conveys a convincing flavor of logic and thoughtfulness – until more details are examined. For example, the Wasp’s routine is to bring the paralyzed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If the cricket is moved a few inches away while the wasp is inside making her preliminary inspection, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is alright. (1963, p. 82)

Apparently, the Sphex will do this repeatedly, no matter how many times one tampers with the cricket. Commenting on the Sphex’s strange behavior, Dennett writes: “Lower animals, such as Sphex, are constitutionally oblivious to many of the reasons that concern them” (1984, p. 24). By reasons Dennett is referring to certain courses of action rationalized by the animal’s own background needs, drives, and (if such states can be legitimately attributed to the animal) beliefs and desires. One problem with the Sphex’s behavior is it appears blind to a wide range of pressing practical reasons, in the sense that the animal can be placed in circumstances that render it systematically poor at achieving its own basic goals.

Now, the range of circumstances in which a system can follow or approximate various behavioral standards will probably vary by degree. In biological creatures, increasingly sophisticated psychological structures correlate with a wider range of behavioral success. And simpler structures correlate with gaps in rational behavior (cf. Hurley, 2003).

For example, the honeybee has evolved a richly combinatorial communicative system – the waggle dance – and a good navigational system. The properties of one honeybee’s waggle dance will tell other honeybees where to go to find nectar. But consider a series of experiments in which Gould and Gould (1988) (and Tautz et al. (2004)) had honeybees discover nectar in the middle of the lake, which they then reported to their colleagues. Almost as if they didn’t believe what they were seeing, the honeybees ignored the waggle dance. One interpretation of this, as Camp (2009) notes, is that the bees put the states <nectar is there> and <there is lake> together into the state <nectar in lake>, which they subsequently rejected. But an alternative interpretation is that the bees failed to make sense of what they saw because of a limit in their representational system. As Camp puts it, “Perhaps their representation nectar there is blocked from interacting with their cognitive map, because the region on the map marked ‘lake’ can’t receive any other markers” (2009, p. 299). If that is right, then the bees have a representational limit that renders them unable to accord with the relevant norm in one circumstance, even though their representational system is overall well-tuned to deliver success.

Like the rest of the animal kingdom, human beings have representational and psychological limitations. But unlike most other animals, human beings have capacities to work with their psychological states and representations in various ways. Penn, Holyoak, and Povinelli have argued that a key feature of the human mind is the ability to reinterpret various kinds of available representations “in terms of higher‐order, role‐governed, inferentially systematic, explicitly structural relations” (2008, p. 127). Tyler Burge is after something similar when he distinguishes between reasoning and critical reasoning.

A non-critical reasoner reasons blind, without appreciating reasons as reasons. Animals and small children reason in this way . . . Not all reasoning by critical reasoners is critical. Much of our reasoning is blind, poorly accessible, and unaware. We change attitudes in rational ways without having much sense of what we are doing. Often, we are poor at saying what our reasoning is. Still, the ability to take rational control of one’s reasoning is crucial in many enterprises – in giving a proof, in thinking through a plan, in constructing a theory, in engaging in debate. For reasoning to be critical, it must sometimes involve actual awareness and review of reasons . . . (2013, p. 74)

At this stage – a stage of adult human sophistication that can involve reflection on our reasons as reasons, and that can involve considerations of relations of reason between our various psychological states, we find a level of agency that Michael Bratman has developed in much detail. We find *planning agents*:

In support of both the cross-temporal and the social organization of our agency, and in ways that are compatible with our cognitive and epistemic limits, we settle on partial and largely future-directed plans. These plans pose problems of means and preliminary steps, filter solutions to those problems, and guide action. As we might say, we are almost always already involved in temporally extended planning agency in which our practical thinking is framed by a background of somewhat settled prior plans. (Bratman 2018, p. 202; see also Bratman, 1999, 2007)

As Bratman’s work makes clear, planning agents are veritably bathed in applicable practical norms. We spend much of our time working through implications of our commitments, testing them against other possible commitments, wondering whether some other course of action might be better in some way, wondering how the plan will impact others, or whether refinements to the plan might make profit along some unforeseen dimension.

There is an interesting series of correlations. As a system capable of behavior (and action) increases in complexity, so do the practical and rational norms that apply to it. So, then, do the chances of internal practical conflict. So, then, does the value to that system of ways of working through, seeking to avoid, seeking to find resolutions to, existing and potential conflict.

It remains to apply this picture of agency to the notion of a disappearing agent, and to draw lessons.

**3. The agent appears**

Although agentive complexity, and arguably internal rational conflict, is present before the agentive sophistication of adult humans, this is the level that motivates Velleman’s and Wallace’s articulation of the disappearing agent problem. We are now in a position to see how this problem afflicts agents. Let us momentarily step back and summarize the discussion.

According to the picture under development, the agent is essentially an integrated system of internal activity and behavioral control that warrants the application of behavioral standards, and contains the capacity or capacities to coherently meet or approximate some sufficient set of these standards. In psychological agents the standards come to be characterized as rational standards, and the activity that leads to coherent behavioral control comes to be characterized as practically rational activity – often, practical reasoning – and action.

At the same time, in psychological agents like humans, the very complexity of the system that is the agent will often lead to internal rational conflict. This is because of the multiple behavioral and rational standards that will apply to such a system in many circumstances. In cases of internal rational conflict, the agent may seem to disappear. This is because in such cases some important features of the agent are at cross-purposes, and if we wish to answer how the agent is implicated in the action, we will struggle to find a good answer.

Now that we have a better handle on what agents are, however, we can see how the problem of the disappearing agent is no real problem, even though it arises from a truth about the nature of some sophisticated agents, like human beings. The problem is no real problem in the sense that it neither challenges any particular causal theory of action, nor does it motivate the posit of esoteric states or causal capacities. The problem is just a function of the ways human behavior and action is produced – namely, by an imperfectly organized set of mechanisms and capacities that sometimes end up at cross-purposes, undermining agential unity or agential rationality. Agents often produce behavior and action that is sub-optimal in one or many respects. This does not make the agent disappear. To think otherwise is to give the agent, qua agent, far too much credit.

**4. Mental action as rational glue**

I have said that as a system capable of behavior (and action) increases in complexity, so do the practical and rational norms that apply to it, as well as the chances of internal practical conflict, as well as the value to that system of ways of working through, seeking to avoid, seeking to find resolutions to, existing and potential conflict.

In human agents, mental actions – imagination, the direction of attention, counterfactual rumination, attempts to remember, etc. – are one of the main ways we have of navigating internal practical conflict. Other philosophers have noted a close connection between mental action and an agent’s capacity to satisfy or otherwise display sensitivity to various applicable norms (Proust, 2013; Metzinger, 2017). The picture in play is one on which many – even if not all – mental actions concern inwardly-directed activities aimed at rationality-relevant states, attitudes, and conditions. So, Thomas Metzinger claims that “Mental action is a specific form of flexible, adaptive task control with proximate goals of an epistemic kind: in consciously drawing conclusions or in guiding attention there is always something the system wants to know, for example the possibility of a consistent propositional representation of some fact, or the optimal level of perceptual precision” (2017, p. 3). When engaging in mental action, an agent is sometimes searching for information, sometimes assessing sets of attitudes for coherence or consistency, sometimes exploring potential consequences of behavior in light of existing beliefs and desires.

So, the picture of agency developed here makes sense of the compelling thought that mental action is tied to human agency (and to agents with similar psychological structure) in an intimate way. I think the pervasiveness of mental action in our mental lives is a product of our particular computational, informational, and cognitive architectural limitations, as well as the solutions evolution seems to have bequeathed to us. I cannot argue the point in full here, but it seems to me that much of our mental action – and especially the actions that contribute to processes of practical deliberation – is driven by uncertainty and conflict (Shepherd, 2015). This uncertainty and conflict are related to our sense of the norms of practical rationality. Often, in deliberation, we are engaged in a search to uncover what it is best to do, or how best to execute a pre-existing intention, or how best to navigate a conflict between various desires, or obligations, or commitments, or whatever. We deliberate because we are informationally fragmented in certain ways (Egan, 2008) – it is a struggle to call to mind the relevant items, and to put them together in rationally satisfying ways.

Velleman and Wallace were heading in this direction, emphasizing various mental actions by which agents more closely approximate the standards of reason. The error is in suggesting that these mental actions should be embedded into the essence of an agent.[[6]](#footnote-6) That suggestion, I have already said, gives human agents too much credit. For a nearly perfect agent may have little need of the mental actions via which humans rationally glue together their many plans and preferences and aspects of identity.

To see what I mean by this, consider a being constitutively incapable of uncertainty or conflict: an omniscient, omnipotent, and fully practically rational being. Call it Al. It is certainly conceivable that Al, in virtue of its supreme knowledge, never faces uncertainty. And Al, in virtue of its full practical rationality, never faces conflict (unless it be a conflict in the very norms of practical rationality). In whatever the situation, no matter how complex, Al need not deliberate to discern the best course of action. We might say that no matter the situation, no matter how complicated or fraught with moral gravity, Al simply sees – takes in at a glance – the thing to do. Al always acquires intentions reflective of Al’s omniscience and full practical rationality. (In order to take in all the information required to always discern the thing to do ‘at a glance,’ Al will need some pretty amazing perceptual sensitivity and some pretty amazing cognitive sophistication. We can assume this is covered by Al’s omnipotence.)

It seems to follow that neither the kind of uncertainty and conflict that is our normal situation, nor the actional processes of deliberation and decision via which we attempt to reduce uncertainty and accord with norms of practical rationality, are essential for agency (Arpaly and Schroeder (2012) and Morton (2017) make the same point). Further, it seems to follow that uncertainty, deliberation and decision are important features of our – that is, human – agency precisely because human agency is far from perfect. We have perceptual, cognitive, and volitional limitations, and it is because of this that uncertainty, deliberation and decision play such a large role in our lives.

Even if these kinds of mental actions are inessential to the nature of agency, for agents like humans, the activity of practical reasoning that is essential to our agency is often conducted via mental actions – intentional mental activities like shifts of attention, inhibition of urges, imagination of possibilities for action or consequences of courses of behavior, comparison of action options, weighing of reasons, and so on. I want to suggest that for human agency, mental action is a rational response to the computational, informational, and architectural limitations we face. Mental action is a kind of rational glue – it is one key way that we attempt to discover the norms of practical rationality, and to enforce rational coherence across the large but disjointed set of goals, preferences, and abilities that we tend to possess.

**References**

Arpaly, N., & Schroeder, T. (2012). Deliberation and acting for reasons. *Philosophical Review, 121*(2), 209–239.

Brand, M. (1984). Intending and acting: Toward a naturalized action theory. Cambridge, MA: MIT Press.

Bratman, M. (1999). Faces of intention: Selected essays on intention and agency. Cambridge: Cambridge University Press.

Bratman, M.E. (2007). Structures of agency: Essays. Oxford: Oxford University Press.

Bratman, M.E. (2018). Planning, time, and self-governance: Essays in practical rationality. Oxford: Oxford University Press.

Brent, M. (2017). Agent causation as a solution to the problem of action. *Canadian Journal of Philosophy, 47*(5), 656–673.

Burge, T. (2009). Primitive agency and natural norms. *Philosophy and Phenomenological Research, 79*(2), 251–278.

Burge, T. (2013). Cognition through understanding: Self-knowledge, interlocution, reasoning, reflection: philosophical essays, Volume 3. Oxford: Oxford University Press.

Camp, E. (2009). Putting thoughts to work: Concepts, systematicity, and stimulus‐ independence. *Philosophy and Phenomenological Research, 78*(2), 275–311.

Clarke, R. (2019). Free will, agent causation, and “disappearing agents”. *Noûs, 53*(1): 76–96.

Davidson, D. (2004). Problems of rationality. Oxford: Oxford University Press.

Dennett, D.C. (1984). Elbow room: The varieties of free will worth wanting. Cambridge, MA: MIT Press.

Egan, A. (2008). Seeing and believing: Perception, belief formation and the divided mind. *Philosophical Studies, 140*(1), 47–63.

Foran, S.T. (1997). Animal movement. University of California, Los Angeles. ProQuest Dissertation Publishing.

Frankfurt, H.G. (1988). The importance of what we care about: Philosophical essays. Cambridge: Cambridge University Press.

Gould, J.L., & Gould, C.G. (1988). The honey bee. New York: W. H. Freeman.

Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language, 18*(3), 231–257.

Keijzer, F. (2013). The Sphex story: How the cognitive sciences kept repeating an old and questionable anecdote. *Philosophical Psychology, 26*(4), 502–519.

Kung, C., & Saimi, Y. (1982). The physiological basis of taxes in Paramecium. *Annual Review of Physiology, 44*(1), 519–534.

List, C., & Pettit, P. (2011). Group agency: The possibility, design, and status of corporate agents. Oxford: Oxford University Press.

Mele, A.R. (1992). Springs of action: Understanding intentional behavior. Oxford: Oxford University Press.

Mele, A.R. (2003). Motivation and agency. Oxford: Oxford University Press.

Metzinger, T. (2017). The problem of mental action: Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 19*. Frankfurt am Main: MIND Group. DOI: 10.15502/9783958573208.

Morton, J.M. (2017). Reasoning under scarcity. *Australasian Journal of Philosophy, 95*(3), 543– 559.

Penn, D.C., Holyoak, K.J., & Povinelli, D.J. (2008). Darwin’s mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*(2), 109–130.

Pereboom, D. (2014). The disappearing agent objection to event-causal libertarianism. *Philosophical Studies, 169*(1), 59–69.

Proust, J. (2013). The philosophy of metacognition: Mental agency and self-awareness. Oxford: Oxford University Press.

Shepherd, J. (2015). Deciding as intentional action: Control over decisions. *Australasian Journal of Philosophy, 93*(2), 335–351.

Shepherd, J. (2021). The shape of agency: Control, action, knowledge, skill. Oxford: Oxford University Press.

Tautz, J., Zhang, S., Spaethe, J., Brockmann, A., Si, A., & Srinivasan, M. (2004). Honeybee odometry: Performance in varying natural terrain. *PLoS Biology, 2*(7), e211.

Velleman, J.D. (1992). What happens when someone acts? *Mind, 101*(403), 461–481.

Wallace, R.J. (1999). Three conceptions of rational agency. *Ethical Theory and Moral Practice, 2*(3), 217–242.

Watson, G. (1975). Free agency. *The Journal of Philosophy, 72*(8), 205–220.

Wooldridge, D.E. (1963). The machinery of the brain. New York: McGraw-Hill.

1. Research for this chapter was funded by the European Research Council’s Horizon 2020 programme, Starting Grant ReConAg 757698. [↑](#footnote-ref-1)
2. This chapter was written over a period of time during which I was also writing a book, *The Shape of Agency* (Shepherd, 2021). Chapter five of that book is devoted to an exploration of the nature of agency. My end-game there is different than here, but both that chapter and this one required development of my thinking regarding the nature of agency. So, the ideas in that book and this chapter bear relations of mutual influence to one another. As a result, what I say in §2 of this chapter extracts, in some cases repeats, and in other cases re-formulates, parts of that chapter. [↑](#footnote-ref-2)
3. The problem of the disappearing agent that I discuss here is thus distinct from Derk Pereboom’s (2014) presentation of a problem for event-causal libertarian views of free will. I mention this because Pereboom calls the problem he discusses the problem of the disappearing agent. See Randolph Clarke (2019) for a penetrating discussion of Pereboom’s argument, and of disappearing agent considerations more generally. [↑](#footnote-ref-3)
4. The dialectic here is a little dirty, since Velleman is not arguing that disappearing agent cases are not cases of action. He is interested, instead, in a notion he calls action par excellence – the exemplification of full-blooded agency. Event-causalist views are not charitably taken as attempting to capture this notion (see Mele, 2003). So, it seems Velleman is better read here as raising a deeper question about the nature of agency, as opposed to a specific problem for event-causal views of intentional action. [↑](#footnote-ref-4)
5. The same need can be seen as motivating earlier proposals – i.e., Frankfurt’s (1988) involving ‘identification’ and Watson’s (1975) involving the agent’s system of values. [↑](#footnote-ref-5)
6. Again, to be fair to Velleman, he embeds this into the essence of agency par excellence (see also Footnote 3). But it is reasonable to be uncertain about the fruitfulness of this category (see Mele, 2003). [↑](#footnote-ref-6)