The Mystery of Agency*

Wolfgang Schwarz

1 Introduction

Humans think a lot about agency: about what people do, what they can do, and what they ought to do. Under philosophical scrutiny, our thoughts lead to difficult puzzles. I'm going to address four of these puzzles. None of them is new, but they are usually discussed in isolation. I believe that they are connected.

My first puzzle involves the "perspectival ought". What an agent ought to do, in the perspectival sense, supervenes on the agent's perspective or evidence. But the perspectival 'ought' also seems to imply 'can'. This suggests that an agent's evidence must settle their ability to perform the relevant act. Couldn't they lack this information?

My second puzzle concerns the description of the available acts. If we want to evaluate an agent's choice from their perspective, their options must be described in a way that does not go beyond what the agent knows. What ensures that the agent has sufficient information to identify their options?

My third puzzle is that we tend to evaluate an agent's options as if they were uncaused "interventions". We consider the consequences of an option but not its causes. But how could someone bring about an uncaused intervention?

This puzzle has a famous sibling. Since we can't influence the laws of nature or the distant past, it seems that we can only have a choice if our actions are undetermined. But we don't view our choices as random chance events either. How could our choices be neither determined nor random?

Together, the puzzles will lead us to a certain picture, a picture in which agents have an infallible and luminous capacity to "will" various things, unconstrained by the circumstances and by the natural order. Explicitly stated, the picture seems crazy. I'll argue that it is, nonetheless, *almost* correct.

2 'Ought' and 'can'

Let's begin with the first puzzle. We often evaluate decision problems from the agent's perspective, bracketing external facts of which the agent may be unaware. For instance, we believe that people should sometimes use maps to find their way around a city, even if it would be better "objectively", in

* Thanks to Mahrad Almotahari, Beri Marušić, Milo Phillips-Brown, Patrick Todd, and the audience at the *Epistemology of Ability* workshop at HU Berlin in 2024 for helpful comments on earlier drafts.

light of all the facts, to take the correct turns without wasting time on consulting the map. Similarly, we think physicians should examine their patients, even though it would be better objectively to skip the examination and immediately start with the optimal treatment. Physicians, gamblers, military commanders, and stock traders have sophisticated rules for how to act given limited information.

When we assess a decision situation from the agent's perspective – wondering what they ought to do, what they ought not to do, what they may do, etc.[1] – we engage in a *perspectival evaluation*, as I'll call it. The defining feature of perspectival evaluations is that they do not draw on facts beyond the agent's perspective: If two possible agents in two decision situations have the same perspective then a perspectival evaluation of their situation should yield the same verdict. This *perspectival supervenience* is the first ingredient in the first puzzle.

The second ingredient is that the perspectival 'ought' seems to imply 'can'. Suppose we ask what an agent – a physician, a tourist, a gambler – should do, given their limited information. We decide that they ought to do so-and-so. This arguably implies, or presupposes, that they *can* do so-and-so. For instance, if the agent goes on to do something else – if they don't do what they ought to do – it would be natural to conclude that they made the wrong choice. But their choice could hardly be wrong if they chose the best of the available options. So the choice we deemed right must have been available.

To get clear on the notion of 'can' that is at issue, let's take a concrete example. Suppose Jones is in the middle of a card game, and we judge that he should play the Ace of Spades. In what sense does this imply that he *can* play the Ace of Spades? It's not enough that he has a "general ability" to play the card, like a pianist who "can" play the piano even if no piano is in sight. No, all the prerequisites for the move must be in place: Jones has the Ace of Spades in his hand, the card is not glued to the other cards, Jones isn't paralysed, his hands are free, and so on. The situation must set so that it is entirely up to him whether he plays the Ace of Spades. Let's call this kind of 'can' – the 'can' that seems to be implied by the perspectival 'ought' – the *practical 'can'*.

Now here's the puzzle. If an agent's perspective settles that they ought to do so-and-so, and 'ought' implies 'can', then their perspective must settle that they can do so-and-so. But couldn't the agent lack information about what they can do (in the practical sense of 'can')?

Here I've assumed that an agent's perspective is related to their information. Intuitively, your perspective comprises what is "given" or "accessible" to you. Perspectival evaluation does not draw on "external" facts that are not given or accessible. These are metaphors. The puzzle arises almost no matter how we flesh them out. A natural idea identifies an agent's perspective with their evidence. More cautiously, I will assume that an agent's perspective settles that they have (or lack) a certain practical ability only if their evidence entails that they have (or lack) the ability.

Suppose, for example, that Jones's evidence leaves open whether he can move his arm. Then his perspective doesn't settle that he can move it: we can imagine an agent with the same perspective who cannot move their arm. This gives us the required connection between perspective and information: if an agent's perspective settles that they ought to do so-and-so, and 'ought' implies 'can', then their

1 I'll keep using 'ought' and 'may', but these words are not important to the point I want to make. Some philosophers have reasoned themselves into a position where they can no longer see that physicians ought to examine new patients. I hope you can at least see that examining the patients is in some important sense "the right thing to do" for a physician with limited information.

evidence must settle that they can do so-and-so.

The puzzle can be strengthened to suggest that the agent's evidence must provide comprehensive information about what they can and cannot do.

For a start, 'ought' doesn't just imply 'can', it plausibly also implies 'can't do better'. Suppose Jones could play the Ace of Hearts, and this would be better than playing the Ace of Spades. Then it could hardly be true that he ought to play the Ace of Spades. Given that the perspectival 'ought' supervenes on the agent's perspective, the agent's perspective – their evidence – must therefore settle not only that they *can* do whatever they ought to do, but also that they *can't* perform various alternatives that would be better if they were available.

What's more, perspectival evaluation can yield judgements not just about what an agent *ought* to do, but also about what they *may* do, what would be *supererogatory*, or what they *should do given that* they are determined not to do so-and-so. In each case, the answer seems to imply that the agent has the practical ability to perform the relevant act. So the agent's evidence must entail that they have this ability.

Finally, our practice of perspectival evaluation has a slot in which we can plug in norms or goals. We can ask what an agent ought to do *by the lights of hedonistic utilitarianism*, or *by Kantian standards*, or *in order to promote their personal wealth*.[2] In each case, answering that the agent ought to do so-and-so seems to imply that they have the required practical ability. Since different norms or goals often select different acts, the agent's evidence must settle that all these acts are available.

In sum, our practice of perspectival evaluation seems to presuppose that agents have detailed information about their practical abilities. This is puzzling as it's clear that people often lack such information.

If an argument leads to a patently false conclusion, we need to reconsider its premises. We might deny that the perspectival 'ought' (etc.) supervenes on the agent's perspective. But this seems to go against the very idea of perspectival evaluation. Perhaps, then, we should deny that the perspectival 'ought' (etc.) implies 'can'?[3] It's not enough for the implication to fail in esoteric cases; it would have to fail almost across the board. This also looks unattractive. Before we jump to either conclusion, let's look at an example.

3 Willings

You may remember Jill the physician, from [Jackson 1991]. Jill's patient John suffers from a minor skin condition. She has a choice between three drugs. Drug A would relieve John's symptoms but not cure his condition. Of drugs B and C, one would cure the condition (without side effects), the other would kill John. Jill doesn't know which of B and C would have which effect, and she can't acquire this information. Let's say that B is the cure and C is the killer. "Objectively", it would be best to give drug B. Given Jill's information, however, the right choice is to give drug.

2 Some reserve the label "perspectival" for evaluations that are tied to the agent's personal goals. The label is not important.

3 I've found this response to the puzzle in [Graham 2011: p.340] and [Way and Whiting 2016: p.1888].

Now suppose that, unbeknownst to Jill, someone has replaced the drug in the packaging of A with drug C, discarding the real drug A. As a result, Jill has false beliefs about what she can do. She believes that she can give drug A, but she can't. What ought she to do? Objectively, giving drug B is still best. But what is the right choice in light of Jill's information? Ought she to give drug A?

I wouldn't say so. What she ought to, I'd say, is something like *giving the drug in the packaging of drug A*. Jill might describe this act as 'giving drug A', but her description would be false.

Information about the unavailability of drug A influences my judgements in a way that information about, say, unknown side-effects of drug A do not. In Jackson's original scenario, where giving drug A is an option, I'm happy to say that Jill ought to give drug A even if, unbeknownst to her, the drug has side-effects that make it better to give no treatment at all. If drug A is not available, by contrast, I want to redescribe the option. (I hope you do, too.) I want to describe it in a way that preserves the link between 'ought' and 'can'. Jill can give the drug in the packaging of drug A, and this is what she ought to do.

Does this mean that we should give up perspectival supervenience? The perspectival 'ought' appears to be sensitive to whether someone has, *unbeknownst to Jill*, tampered with the drugs. In Jackson's original scenario, Jill ought to give drug A. In my modified scenario, she instead ought to give the drug in the packaging of drug A. Yet her perspective remains the same.

One might, however, put pressure on the judgement about the original scenario. Focus on the original scenario. Nobody has tampered with the drugs. But this isn't entailed by Jill's perspective. (Hence the threat to perspectival supervenience.) Her perspective is compatible with my modified scenario, or with scenarios in which drugs A and C have been swapped. These possibilities are *unlikely* in light of Jill's perspective or evidence, but they are not impossible. Let's say that there's an 0.01% probability that drugs A and C have been swapped, given Jill's evidence. As a consequence, Jill's evidence doesn't settle how she would have to go about giving drug A, which packaging she would have to open. Now reconsider the question what she ought to do. Ought she to give drug A? I'm reluctant to say so. I'd say she ought to give the drug in the packaging of drug A, of which she can be almost (but not entirely) certain that it is drug A.[4]

A third answer to our puzzle now comes into view. We can accommodate ignorance of one's practical abilities without denying either perspectival supervenience or the link between normative judgements and practical ability. If an agent doesn't know whether they can do so-and-so, *doing so-and-so* is not a suitable object of perspectival evaluation: it isn't a candidate for something the agent ought to do. On closer inspection, the premises of our puzzle don't entail that agents must have comprehensive information about their practical abilities. It is enough that anything an agent can do is identifiable by the agent under a description (like, perhaps, *give the drug in the packaging of drug A*) which they know they can make true.

In a real-life scenario, Jill's evidence probably wouldn't settle that she can give the drug in the packaging of drug A. Couldn't the packaging be empty? Couldn't the security forces, or the laws of nature, prevent her from opening the packaging? Couldn't she be about to have a heart attack, or lose control of her arms? These possibilities may be far-fetched, but any of them might obtain while her

4 I assume that the relevant notion of evidence is factive. On a non-factive construal of evidence, it would be implausible that an agent's perspective settles that they can do so-and-so iff their evidence entails that they can do so-and-so.

perspective remains the same.

Suppose there is a small but positive probability, given Jill's evidence, that the muscles in her arms have just stopped working. What should she do? What is the right choice, in light of her information? As before, it doesn't seem right to say that she ought to do something of which she is rationally unsure whether she can do it. Strictly speaking, what she ought to do is not *open the packaging of drug A* (etc.), but *try to open the packaging* (etc.).

We are led to the conclusion that the options that figure in perspectival evaluation are usually not overt acts but mental acts of *trying*, *intending*, *deciding* or *willing* whose practical availability is not in doubt.[5]

You may wonder how an agent could know which of these mental acts they can perform. Can't we be unsure about what we can try or will? I'll come back to this worry. You may also object that our practice of perspectival evaluation doesn't align with the conclusion I've reached. We rarely talk about what agents should try or will. We say that Jones should play the Ace of Spades, or that a physician should examine their patient. Jackson says that Jill ought to give drug A. I want to briefly address this objection.

When we evaluate an agent's choice, we normally take for granted certain facts about the connection between their narrowly construed options – the available "willings", as I'll say from now on – and overt acts, even if these facts are not strictly entailed by the agent's perspective or evidence. We ignore the possibility that the agent might have a heart attack, or lose control over their arms. If these possibilities are set aside, the agent's evidence does entail that they can perform ordinary, overt acts. When we think about Jackson's scenario, we also ignore the possibility that someone might have swapped around the drugs in their packaging. Jill's evidence then entails that she gives drug A in every non-ignored world in which she wills to give the drug in the packaging of drug A.

Now suppose the perspectival modals 'ought', 'should', 'may', 'can', etc. are context-dependent in roughly the way [Lewis 1996] argued that 'know' is context-dependent. The specifics depend on independent questions about how these expressions work, but the basic idea is simply to restrict all quantification over worlds to worlds that aren't "properly ignored" in the conversational context. For instance, a conditional analysis of the practical 'can' might say that 'S can ϕ' is true iff S ϕs in the "closest" non-ignored worlds in which S wills to ϕ.[6] Similarly, if we start with the hypothesis that 'S ought to ϕ' is true (on its perspectival reading) iff S's evidence entails that the closest worlds in which S ϕs are at least as good (by the relevant standards) as the closest worlds in which S performs any other act that S can perform, then on the revised approach, we would restrict both the evidence and the relevant closest worlds to ones that aren't properly ignored.

Setting aside the details, the idea is that 'Jill ought to give drug A' is true in Jackson's original scenario, as long as we ignore far-fetched possibilities. It becomes false if attention is drawn to these possibilities. It is also false in my modified scenario, where one of the far-fetched possibilities is

5 This view has been defended, on more or less the present grounds, in [Ross 1939], [Prichard 1949], [Zimmerman 2008: ch.3], and [Hedden 2012].

6 I prefer a slightly different analysis: S can ϕ iff there is an accessible world in which S ϕs as the result of a willing X such that S's evidence, minus any properly ignored worlds, makes it certain that S ϕs conditional on X, where a world is *accessible* iff (roughly) there is a willing Y such that the world is amongst the closest not-properly-ignored Y worlds. See [Schwarz 2020].

actual: the actual world is never properly ignored.

Let me wrap up. We began with a puzzle: an apparent tension between perspectival supervenience and 'ought'-implies-'can'. The best response, I suggested, is not to give up one of these assumptions, but to posit a special domain of "willings" as ultimate objects of perspectival evaluation, along with a contextualist account of agentive modals that explains why we rarely talk about what agents should will. There remains a puzzle about the newly introduced willings. What are they? Why would agents always have perfect information about which of them they can perform?

4 Expectations

On to the second puzzle. (I'll be quicker from now on.) I have said little about how the perspectival evaluation of an agent's options works. How do we assess what an agent should do, given their evidence or perspective? A standard answer is to look at the possible outcomes (broadly understood) of each option and adding up their value, perhaps weighted by their probability given the evidence. Decision theory makes this precise, but the details aren't important. What matters is that this approach only yields sensible results if the agent's options are individuated in a certain way.

Back to Jackson's case of Jill and John. Jill knows that one of drugs B and C would cure John and the other would kill him. Since drug B is the actual cure, an act of *giving drug B* can also be described as an act of *curing John's ailment*. But this is not an adequate description for assessing Jill's options. If Jill had an option of curing John's ailment (without side effects), this would obviously be the best choice. An act of *curing John's ailment* is certain to have a good outcome. We couldn't say that Jill ought to give (or will to give) drug A.

The lesson is that an adequate description of an agent's options must not draw on facts outside the agent's evidence. The options must be described in a way that is *transparent* to the agent, we might say. What exactly does this require?

It's not enough that the agent knows they can perform the so-described act. In Jackson's story, Jill knows that one of her options amounts to curing John's ailment. She knows that she can cure John. But she doesn't know how.

An adequate option description must leave no room for such ignorance. If ϕ is an option, the agent must have no rational doubt about which choice would amount to ϕing. As a corollary, the agent should be in a position to become rationally certain that they are ϕing merely by making a decision.

Clearly, overt acts usually don't meet this condition. In a real situation, Jill could hardly be certain that she will be giving drug A merely on the basis of making a decision. As we noted before, her evidence doesn't settle that the packaging contains the drug or that her arms will keep working. A full evaluation of her decision problem should take these possibilities into account. If there is a small but positive probability, given Jill's evidence, that drug A has been replaced with a lethal alternative, then *giving drug A* is not an adequate description of the option, as it would rule out the possible outcome of killing John.

For any ordinary act, there is a possibility of deciding to perform it and then failing to perform

7 That realistic options should be construed as "acts of the will" is a popular view among decision theorists, for the reason I have given. See, for example [Sobel 1983], [Weirich 1983], [Joyce 1999: ch.2].

it, and this possibility is not conclusively ruled out by the agent's evidence. The transparency requirement therefore suggests that the options shouldn't be described as ordinary acts, but as peculiar internal acts of trying or willing for which no such possibility arises.[7]

The conclusion we've reached matches the conclusion from the first puzzle. But the argument is different. The earlier argument relied on a connection between the perspectival 'ought' and the practical 'can'. In the present argument, the sole premise is the transparency of options.

As before, one might worry that even willings or tryings don't meet the transparency condition, that there is always a gap between deciding to ϕ and ϕing, no matter what ϕ might be. But then how are we to evaluate a decision situation from the agent's perspective? Standard formulations of decision theory all assume the transparency condition, and no viable alternative is in sight, as I've argued in [Schwarz 2021].

5 Interventions

My next puzzle turns on another aspect of how we evaluate an agent's options. Suppose an agent has a choice between X and Y. Let's not worry about whether these are willings or ordinary acts. To figure out what the agent ought to do, we examine what might happen *as a result* of choosing X or Y. We reason forward from the hypothetical choice, not backwards.

Suppose, for instance, that poor Jack needs to go to town. He could take the bus for £2, as usual, or he could hire a limousine with a personal driver for £2,000. something he never does because he can't afford it. He would hire the limousine only if he were immensely rich. Yet, we don't reason that Jack should take the limousine because he would be immensely rich if he did. This would be reasoning backwards, from the hypothetical choice to earlier circumstances that might have brought it about. We don't do this.

We evaluate an agent's options as *interventions* that affect the world's causal structure "from the outside", without explicable causes.[8] The limousine option is bad, we think, because it would be a waste of Jack's scarce money. We don't ask what might have led Jack to choose the option. This is puzzling, for we also know that an agent's choices have causes. They are not uncaused interventions.

We can sharpen the puzzle. Suppose Jocasta believes that the history of the universe follows certain deterministic laws L. Today, she is asked if it does. She answers 'yes'. Assuming her only relevant aim is to tell the truth, this is intuitively the right choice. She is confident that answering 'yes' will achieve her aim. But what would have happened if she had answered 'no'? Or rather, what do her beliefs imply about what would have happened?

Let w be an arbitrary world of the kind Jocasta believes she inhabits: a world that conforms to L and where she answers 'yes'. Since L is deterministic, any world in which she answers 'no' either doesn't conform to L or differs from w in its entire history. (This follows from standard definitions of determinism.) By the criteria of [Lewis 1979], the "closest" worlds to w at which Jocasta answers 'no' are worlds that don't conform to L. According to Lewis, Jocasta would have answered truly if he

8 The hypothesis that an agent's options should be evaluated as uncaused interventions, at least from the agent's own perspective, has a venerable history. See, for example, [Kant 1781: A542/B570–A557/B585], [Ramsey 1931], [Ismael 2007], [Solomon 2021].

had said 'no'. We don't get the intuitive result that Jocasta was right to answer 'yes'.[9]

[Dorr 2016] and others have argued that the relevant closeness measure should hold fixed the laws and vary the past: if Jocasta had answered 'no', they hold, the world would still have conformed to L but its entire history would have been different. But suppose Jocasta believes that some event E took place in the past, and she is asked whether it took place. She answers 'yes'. If E involves an aspect of the past that would have been different if she had said 'no', it follows that Jocasta would have said something true, and we don't get the intuitive result that answering 'yes' was the right choice.[10] And why couldn't E describe some such aspect of the past?[11]

The problem is that neither the distant past nor the laws of nature are intuitively a *result* of Jocasta's present choice. In a deterministic world where Jocasta answers 'yes', the laws and the past are preconditions that *lead* to her choice. Since our practice of perspectival evaluation requires reasoning forward from the hypothetical choice and not backwards, we want to hold fixed the laws and the past when we evaluate Jocasta's options. But this looks impossible. If we hold fixed the laws and the past, and the laws are deterministic, how can we vary Jocasta's choice?

The present puzzle is related to a fourth one, canonized by [van Inwagen 1983]. The fourth puzzle starts with the plausible assumption that the distant past and the laws are outside our control: there is nothing ordinary people can do that could make any difference to the past or the laws. For any choice we can make today, if we were to make it, the distant past and the laws would be just as they actually are. If the laws are deterministic, it seems to follow that the only choices we *can* make are the ones we *will* make.

One can't escape this puzzle through a compatibilist view of freedom. The puzzle is not so much about freedom, but about practical ability, or choice – assuming that we don't have a choice if we only have one option. Given our practical inability to affect the past or the laws, it seems that we have no practical abilities to do otherwise, and no choices to make, unless the world is indeterministic. Unfortunately, it is unclear how indeterminism would help. Some versions of quantum physics suggest that the dynamics of our world is thoroughly random, subject only to statistical regularities. If our choices are random processes, having more than one option does not require a power to influence the past or the laws. But we don't think of our choices as random events. When we wonder what an agent should do, we don't wonder how some chance process that determines the agent's behaviour should go. The normative 'should' does not even seem appropriate if we're talking about the outcome of a chance process. We take for granted that the agent has a kind of control over their choice that is hard to reconcile with the idea that the choice is entirely random.

9 Worse, suppose Jocasta is not 100% certain that the world conforms to L. In worlds where L is false, his utterance of 'yes' does not achieve his aim of telling the truth, but an utterance of 'no' would almost certainly have done so. In worlds where L is true, both utterances achieve his aim. On balance (in expectation), we have to say that Jocasta should have said 'no'! See [Ahmed 2013].

10 As in the previous footnote, we can easily get the opposite result that Jocasta should have answered 'no', if Jocasta isn't 100% certain about E.

11 On Dorr's view, it would be hard for Jocasta to have evidence (in worlds like ours) for an aspect of the past that depends on her present choice. But we can evaluate her choice in light of her beliefs, without assuming that her beliefs are based on evidence. One might also suggest that the respect in which the past would be different depends on what Jocasta is asked about today, so that the past would always be different in some respect other than E. This seems mysterious. It also doesn't help if E happens to contain enough information about the past so that it, together with the hypothesis that the history conforms to L, entails that Jocasta answers 'yes'.

We are led to the abstruse conclusion that an agent's choices are neither determined nor random. Somehow, they impinge on the physical world from the outside, from a special realm that is not subject to natural laws.[12]

6 The picture

Let's review the picture that has emerged from the four puzzles. In sections 2–4 we met some arguments suggesting that an agent must have perfect information about their options. Since people generally don't have perfect information about which ordinary acts they can perform, this led us to posit special acts of "willing" whose availability is never in doubt. The considerations in section 4 further suggested that willings are special in that there is no epistemic gap between deciding to act and acting: if an agent decides in favour of a willing, they can be rationally certain that they succeed.

I use 'willing' as a technical term, as a placeholder for whatever these special acts might be. Some authors prefer 'trying' or 'intending' or 'deciding'. The label is not important. We shouldn't assume that any word of ordinary English unambiguously captures the target.

Willings don't seem to be individuated physically or functionally. For any non-trivial physical or functional condition, one can easily imagine an agent who is rationally unsure whether they can bring about an event that satisfies the condition. Willings seem to inhabit a different kind of space, a luminous realm that is directly and infallibly accessible to the agent.

The other two puzzles, discussed in section 5, pointed in the same direction. They suggested that the acts an agent can choose should be understood as "interventions" into the physical world from a realm that is not subject to the natural laws.

This is the picture that has emerged. I'm not saying that it is correct. All I'm saying is that something like this picture appears to be implicit in our thinking about agency. Each of the puzzles began with assumptions that seem plausible, that seem to be part of our conception of agency. The picture is what we get when we follow these assumptions.

Some may find the picture appealing. To me, it looks crazy. Nothing in current science suggests that human agency involves a special realm of non-physical events. Even if there were such a realm, why would agents have perfect information about it? Where would that information come from? Is it a priori? Revealed by a special, infallible kind of perception? By an infallible intuition? No answer looks remotely plausible. The interventionist aspect of the picture looks no better. Real choices are not uncaused interventions. They have a causal explanation. Jack takes the bus rather than the limousine because he is aware of his pecuniary situation and doesn't want to bankrupt himself. His choice is fully explained by his goals and information. For all I know, it may also be determined by the distant past and the laws: determinism is a live possibility in contemporary physics. At the very least, I can conceive of agents like Jocasta (from the previous section) who rationally *believe* that their world is deterministic. If we evaluate options as undetermined interventions, every possible choice entails the falsity of determinism: We reach the bizarre conclusion that every rational agent should deny determinism, irrespective of their evidence.

I'm not sure *how* crazy the picture is: whether it is merely false or downright incoherent ("not even

12 Many renditions of this idea have been explored in the literature on free will; see [Clarke et al. 2021] for a survey.

false"[13]).

It's tempting to think that the epistemic side of the picture is refuted by the conceivability of Frankfurt cases. Suppose an agent's deliberation process is monitored by a demon determined to prevent a particular choice if they see it coming. This seems to make the choice practically unavailable. You can't even will to raise your arm if a demon would prevent you from doing so. But you needn't know that the demon is there. If an agent has inconclusive information about whether they are in a Frankfurt case – and don't we all, if we put on our philosophy hats? – they have inconclusive information about their options, even if the options are willings.

The point could also be made with phobias. Ever since [Lehrer 1968], It is widely accepted that agents with severe phobias can't even *try* or *intend* or *will* to perform certain acts that would otherwise be available. An arachnophobe can't even will to touch a spider. Presumably, however, an agent may have inconclusive information about whether they suffer from arachnophobia. If so, they have inconclusive information about the available willings.

Here, however, the causal side of the picture might come to the rescue. If willings are instantaneous and uncaused interventions, perhaps no demon could "see them coming". If an agent represents their options as uncaused interventions, they might conclusively rule out being in a Frankfurt case.

The same is true for phobias, if these are assumed to work like the Frankfurtian demon: that is, if they are assumed to prevent the relevant choice (by causing the agent to faint or have a panic attack) shortly before or while it occurs. If willings are instantaneous and uncaused interventions, no psychological process could foresee and prevent them or stop them in their tracks.[14]

Personally, I'm not convinced by the "Frankfurt model" of phobias. Phobias are a type of fear. Like other fears, they arguably work by motivating the agent to avoid certain acts. You walk carefully on the ice because you fear that it might crack. Your fear doesn't take away the option to stomp. Proof: If the incentives were right, you would stomp, despite your fear. Fear merely marks the relevant option as undesirable. Extreme phobias might be impossible to overcome, no matter the incentives. But they might still function in the same way. In decision-theoretic terms, they would work not by removing options, but by associating them with very low utility. (Compare [Alvarez 2013] on addiction.) On this model, an arachnophobe *can* will to touch a spider, in the practical sense of 'can', and cases of unknown phobia don't pose a problem for the picture to begin with.[15]

If the picture is incoherent, I suspect the incoherence lies in not in its epistemic, but in its causal element, in the idea of rational acts that are neither determined nor random. "Libertarians" about free will have filled many volumes trying to explain how this might be possible. I am not convinced that

13 "Nicht einmal falsch" (Wolfgang Pauli), often translated as "not even wrong".

14 We have to be careful about timing. The conclusion we've reached from the first two puzzles is that an agent has perfect information about what willings they can realize right now. They need not have perfect information about what they can do in the future, if only because they may not be sure whether they'll still exist. The puzzle from section 4 may appear to suggest that a deliberating agent must have perfect information about their *future* options, about what they can do at the end of the deliberation. But this isn't so. It suffices that the agent is sure about what they can eventually choose conditional on their continued survival and agency.

15 Why do people intuit that arachnophobes can't will (or intend or try) to touch a spider? Perhaps they tacitly assume the Frankfurt model of phobia. Or perhaps they focus on an "impure" sense of 'can' that holds fixed the satisfaction of relevant norms or desires – the sense in which one "can't" come to a party because one has other plans. Or perhaps they have in mind extended activities of trying (and willing?) that could be intercepted by panic attacks.

they have succeeded. The task is especially challenging for the picture I have described because it posits a libertarian kind of control even for cases like Jack's, where all the reasons point in favour of one option and against the other. Jack *could* take the limo. He takes the bus not because it's the only option but because the alternative is a foolish waste of money. How can we treat him as a rational agent and yet assume that his reasons, up to the point of decision, somehow leave open what he will choose?[16]

7 Armchair robotics

Let's take a short break and think about how we might design a robot. (Bear with me. The point of the exercise will become clear soon.) The robot we are designing has a central processor and a database to store goals and information about the world. The robot also has a motor system that can be activated by the central processor. Picture wires running from the robot's wheels and limbs to the central processor. Depending on the electrical current on the wires, the wheels and limbs move in different ways. We have to design an algorithm that determines which electrical signal is sent down which wires, based on the contents of the robot's database.

Here's one approach we could take. First, we figure out what movement each signal causes. Suppose activating the red wire normally causes the robot's left arm to rise. Whenever the database indicates that raising the left arm would bring the robot closer to its goals, we let the motor interface activate the red wire. Similarly for the other signals.

This would work if activating the red wire always causes the left arm to rise. However, imagine the robot is positioned next to a wall, so that it can't raise its left arm. The signal in the red wire would now cause the arm to press against the wall. This might be a good idea, or it might be a terrible idea, depending on what the robot wants to achieve and what it knows about the wall. We want to send the signal if *pressing against the wall* is expected to lead to good outcomes.

This kind of case is common. Depending on the incline of the floor and the direction and speed of the wind, the same motor signal might cause the robot to accelerate, maintain pace, or slow down. Worse, the robot may not have perfect information about which of these conditions obtain, and therefore about what would happen if a given signal were sent.

Here's a better approach. We introduce primitive elements $X1, X2, X3, \ldots$ into the robot's database and decision algorithm. I'll call them *motor commands*. The motor commands serve as the options from which the robot chooses. We design the robot so that the red wire is activated iff the robot chooses $X1$, the blue wire iff it chooses $X2$, and so on. To the robot's database, we add information about what is likely to happen as the result of the motor commands: $X1$ normally causes the robot to move forward, $X2$ causes its left arm to go up, and so on. We can also add information about how these outcomes depend on the circumstances: if the robot's left arm is obstructed, $X2$ causes the arm to exert force against the obstruction. Motor commands are not descriptions of real acts. $X2$

16 Libertarians often restrict their account to options between which the agent is torn, and sometimes locate the relevant "intervention" at a time before the agent's choice. This may help to make the account more plausible, but it isn't compatible with the picture I have described. My route to libertarianism is, of course, non-standard. I have mentioned neither free will nor moral responsibility. My topic is how we evaluate an agent's choices, not the preconditions for moral blame or praise. The topics are related, but not the same.

does not mean *I lift the left arm*, *I press against the wall*, or *I send current down the red wire*. It is a primitive element in the robot's database and decision algorithm, logically independent (in the robot's doxastic space) of any real acts or event. Its purpose is not to represent an aspect of reality, but to allow for a simple and versatile interface between the robot's information and goals on the one hand and its motor system on the other.

In normal situations, the robot may be confident about what would happen if it chooses $X2$. If there is no reason to think that the left arm is obstructed, we (and the robot) might describe a choice of $X2$ as a choice to lift the left arm. Alternatively, we describe $X2$ as sending a current down the red wire. If the robot happens to know about its inner workings, it might also think of $X2$ this way. But our design does not require such knowledge, and it does not require a perfect association between motor commands and electrical signals: our design still works if a choice of $X2$ doesn't lead to activation of the red wire if the robot's battery is low or the temperature extreme.

This is – roughly – how I would design a robot. It is – roughly – how actual robots are designed. I suspect it is – roughly – how evolution has designed us.

Now imagine an agent who works like this. Call him J. At the moment, J isn't sure whether his left arm is obstructed. If it isn't, lifting it would be a good idea given his goals and information. If the arm is obstructed, it would be better to move back a few steps, but it wouldn't be a disaster if he tried to lift it. What should he do?

We may not want to say that he should lift his arm. He certainly wouldn't evaluate the relevant option as *lifting the arm*, and such an evaluation would yield the wrong results. (It would disregard the possibility that the arm is obstructed). J might decide that $X2$ is his best option, better than $X1$ or $X3$, etc. But how should *we* think about his choice? We have no idea how he represents his options, and it wouldn't help us if we did: '$X2$' means nothing to us.

What we can say is that J should choose *whichever motor variable he expects to lift his arm*, or something like that. We might introduce a shorthand: J should *try to lift his arm*, or *decide or will to lift his arm*. On this understanding, 'trying' or 'deciding' or 'willing' doesn't describe a special type of act. It hints at something only the agent can properly represent to themselves.

Even J, however, need not think of his options as motor commands. If $X2$ normally causes his arm to go up, and is likely to do so now, he, too, may conceptualize $X2$ as trying (or willing, or deciding) to lift his arm.

As J goes through life, he is often unsure about which overt acts he can perform. He doesn't know if his arm is obstructed, if he has enough energy, if the floor is slippery. He takes the different possibilities into account when he chooses a motor variable. Would he also need to consider if he can realize a particular motor command? If he can "will" to lift his arm? Perhaps not.

What would a situation look like in which J can't realize, say, $X2$? From an objective point of view, there are no such situations, as '$X2$' doesn't designate a real kind of act (or event or proposition). There are situations in which J can't send the electrical signal associated with $X2$, because he is turned off. But that's not a possibility he needs to take into account when making a choice. Other situations in which he can't send the signal are Frankfurt cases where the relevant choice would be anticipated and prevented. One might argue that J should treat these as scenarios in which he can't realize $X2$. But Frankfurt cases are not only unusual and strange, it's also not clear what an agent

should do if they suspect they are in such a case. Suppose J would like to lift his arm, by choosing $X2$, but suspects a demon would prevent the choice. What should he do? One is tempted to say that he should *try to choose* $X2$, and that he should take the possibility of failure into account. But this would only introduce another layer of options, where the same problem would arise: What should he do if he suspects that a demon would prevent the decision to try to choose $X2$?

The upshot is that an agent like J has little use for a capacity to be unsure about the available motor commands. He may take it for granted that all his motor commands are on the table. He may be designed so that he is always certain about which "willings" he can perform.

Now let's think about how J might evaluate his options. What does he believe would happen if he chose $X2$? If his cognitive system works as I've suggested, his database will indicate what might happen as a result of choosing $X2$. It might tell him, as a general rule, that his arm will tend to go up unless it is obstructed. It might also tell him that his arm is currently not obstructed, so that J can infer that his arm would probably go up. But would his database tell him how $X2$ would have come about? Perhaps not.

Different considerations support letting his database fall silent on this matter. Some have to do with computational economy and with physical asymmetries in worlds like ours. I want to focus on a simpler point. We don't want J to "reason backwards" when he evaluates his options. Suppose he believes that if he were to choose $X2$ then the past would be different in a certain respect, and he lets this belief guide his choice. He might then opt for $X2$ merely because he prefers the alternative past associated with $X2$. We wouldn't want our robot to reason this way, given that it has no real influence on the past. For the same reason, we wouldn't want our robot to think that different choices come with different laws of nature.

J's cognitive system may represent his motor commands as *interventions*, with (more or less) predictable consequences, but no predictable causes. It wouldn't thereby represent the motor commands as random – as outcomes of a well-defined chance process. No, nothing at all would be said about their origin. They would seem to interfere into the world's causal structure from the outside, from an apparent realm where causal questions don't arise.

I'm not suggesting that an agent *has* to be designed in this fashion. There might be clever ways to deal with uncertainty about Frankfurt cases, or to prevent backward reasoning without representing choices as interventions. But the alternatives aren't obvious, and they tend to be complex. The design I have described, with its luminous realm of interventionist options, is comparatively simple. I'm suggesting is that something like this design might be implicit in our conception of agency.

8 Unravelling the mystery

In the first half of this paper, I have argued that a curious picture seems to lie behind our conception of agency, a picture in which the immediate objects of choice are special acts ("willings") that intervene into the world's causal structure from the outside and whose availability is never in doubt. The picture looks indefensible. In the previous section, we saw that it may nonetheless be largely correct – provided that our picture of agency involves agents with a certain cognitive architecture.

A caveat. The architecture I have described is severely idealised. Real humans have control over

a fine-grained and high-dimensional space of body movements and an equally rich space of mental acts, including the formation of commitments and plans. An efficient handling of this richness calls for a more sophisticated design. I won't enter into these complications, except to note that they might explain some respects in which the picture I have described is at odds with our experience.

In conscious deliberation, we rarely think of our options as willings or tryings, or even as body movements. In a restaurant, I might deliberate between *having pizza* or *pasta*. Having chosen the pasta, I might choose a string of words to express my decision, but this is typically sub-personal, as is the choice of muscle movements in my tongue and lips that produce the utterance. My decision to have pasta makes me confident that I will indeed be having pasta. But I recognize the possibility of failure: the restaurant could be out of pasta, the waiter could refuse to serve us, my muscles could fail to obey me. I could have accounted for these possibilities in my deliberation: I could have asked whether I should *try to have pizza*. But the costs of the failure possibility are essentially the same for pizza and for pasta, so nothing is lost by ignoring them. I could also have asked whether I should *try to have pizza by uttering such-and-such words* or even *by moving my mouth in such-and-such manner*. But it's easy to see why our cognitive system would delegate these choices to "lower-level" systems: the possible costs and benefits of the realistic options are essentially on a par, and don't affect whether I should have pizza or pasta. (It might be otherwise if the pasta had a long and difficult-to-pronounce name and I cared about not embarrassing myself.)

Let's return to the simple, non-hierarchical picture. Is it true of J, whose decision system directly deliberates over motor commands? Does J have the practical ability to perform special acts that intervene in the world's causal structure and whose availability is infallibly known to J? Of course not. The picture is not true of us, not of J, and perhaps not of any possible agent. Can we tidy it up? Can we fix our picture of agency so as to make it correct?

It's not easy. The willings in the centre of the picture are fictions. When J "chooses $X2$", or "wills to lift his arm", what happens in reality is that his decision module sends an electrical signal down the red wire and updates his database so that $X2$ becomes certain. Neither we nor J could sensibly replace the fictitious willings in our reasoning about his choices with any such real events. We can't practically do it, for lack of physiological knowledge. But we also can't do it unless we want to drastically revise our picture of agency. Electrochemical events in an agent's decision system are neither uncaused interventions nor is their availability directly accessible to the agent. We would need a radically different model.

I don't know if there is an alternative to the picture I have described that is both workable and correct. I suspect there is none. The only correct alternative would represent agents as physical systems. We would gain in empirical accuracy, at a great loss in tractability. Perhaps the best we can do is to play along with the fiction. We'll run into puzzles. We may wonder how an agent could have infallible information about their options, or how to make sense of rational but uncaused interventions. We may be stumped by cases like Jocasta's. But if we understand that the fiction is a fiction, perhaps we needn't be too worried about the puzzles it raises.

9 Postscript: The epistemology of ability

Let me say a little more on the topic of this volume: the epistemology of ability. In the picture I've sketched, agents have perfect knowledge about the available motor commands. No empirical or non-empirical basis is needed for this knowledge. What about knowledge of more ordinary abilities? How could J know that he can, say, raise his left arm?

The question has many readings, if only due to the polysemy of 'can'. Let's focus, as a beginning, on the "specific", "transparent", and "pure" reading, where we're concerned with what J can do foreseeably, in the present circumstances, even if it might violate his inclinations or relevant norms.[17] In essence, our question is: how could J know that choosing $X2$ would raise his arm? More colloquially, how could he know that he would succeed if he tried (or decided, or willed) to raise his arm?

I'm interested in a particular kind of knowledge. To illustrate what I have in mind, note that one can challenge J's knowledge by pointing out that his left arm might have been amputated last night, and that he might now be hallucinating the arm. If he doesn't know that he has a left arm, he plausibly can't know that $X2$ would raise the arm. In response, one might claim that J *sees* that he has a left arm, and that seeing is a kind of knowing. This may be right, but the skeptic still has a point. Suppose before J opened his eyes today, he had reason to suspect that his arm has been amputated and that he would have a realistic hallucination of it. Suppose he rationally gave probability 0.8 to this hypothesis. How should this assessment change when he opened his eyes and (as we might put it) saw his arm? Answer: it should remain unchanged at 0.8. In general, J's visual experience does not affect his rational credence in the amputation/hallucination hypothesis. If his prior credence in the hypothesis was 0.00001 (as it might well have been if he had no special reason to take the hypothesis seriously), his posterior credence should still be 0.00001. In that sense, the skeptic is right: J cannot rule out the amputation/hallucination hypothesis, even if he can see that he has arms. What we should say, against the skeptic, is that not all open possibilities are on a par: absent special reasons, J may rationally give negligible credence to the amputation/hallucination hypothesis.

Let's now turn to our main topic. Granting that he has arms, how could J know that his left arm would go up if he tried to raise it, if he chose $X2$? There are no a priori connections, in J's database, between motor commands and ordinary events. As a result, there are countless error possibilities. Some are obvious: the arm would not go up if it is tied down or too close to a wall. But even if all the usual prerequisites are in place, there's no guarantee that the arm will go up. How can J rule out all these error possibilities?

Could affordance perception do the job? Some hold that we perceive the objects in our environment as affording certain actions: we see a door knob as *grabbable (by us)*. Assuming that something is grabbable by an agent only if the agent can grab it, it would follow that we can see that we can grab the door knob. One might also posit a special internal sense of our body and the actions it affords. Either kind of perception might tell J that he can raise his arm. (Maier, this volume, is sympathetic to these ideas.)

As before, I don't think this gets to the heart of the challenge. Suppose you have just undergone surgery, and you are rationally unsure whether you can raise your arms. Your internal sense of what

17 See [Maier 2018] for the distinction between "specific" and "general", [Schwarz 2020] for the distinction between "transparent" and "effective", and [Kratzer 1977] for that between "pure" and "impure", though not with these labels.

your body affords can't tell the difference, nor can your eyes: the door knob would look just as grabbable either way. Affordance perception should not affect your credence in whether the surgery was a success. In that sense, it provides no information about your ability. The same is true for J. If he woke up with credence 0.0001 in any of the error scenarios, any affordance perception he might have should leave his credence unchanged.

I don't deny the reality of affordance perception. It may well be that our visual system somehow directly "tells us" that an object is grabbable, more or less as it "tells us" that we have arms. But I deny that we can rationally take the testimony of our senses as a direct foundation for our beliefs. In Bayesian terms, I deny that we may rationally conditionalize on what our senses tell us. The psychological story may be simple and direct, the epistemic story is not. Here I agree with Schoonen [?] and Bruineburg, this volume, although I suspect that my version of the epistemic story is different from theirs.

The epistemic story I favour is the standard Bayesian account of induction. When I introduced motor commands, I suggested that our robot's database should specify what might happen as a result of the various commands: that $X2$ tends to raise the agent's left arm, except if a wall is in the way, in which case it would cause the arm to press against the wall, etc. We might have explicitly programmed all this information into the robot's database. But this is cumbersome, inflexible, and easy to get wrong. Much better to let the agent discover the associations between motor commands and ordinary events.

The most basic method of discovery is trial and error. J might have been built so that he randomly chose among the available motor commands in his earliest weeks, observing their effects. He could thereby have observed that $X2$ tends to move his left arm, and not his right arm or the nearby table, unless these are connected to his left arm. He could have observed that $X2$ does not move his arm through walls and other barriers, and so on.

There is no logical guarantee that past associations will hold in the future. Accordingly, J should always reserve some credence for the possibility that $X2$ won't lead to the left arm going up next time, even if there's no wall in the way, etc. But he needn't worry too much about this possibility. Rational credence can (and should) be biased towards worlds in which past regularities persist into the future.

When J has observed his arm going up after issuing $X2$, he initially learns about a "specific" (and "effective") ability: that he was able to raise his arm on this occasion. The inductive generalisation gives him knowledge of a "general" ability: that he is able to raise his arm under such-and-such conditions. Knowing that the right conditions obtain, he can infer that he has the "specific" (and "transparent") ability even if the ability isn't exercised.

There are many other ways to learn about one's abilities. We can, for example, generalize from observed abilities to different abilities. If J has never lifted an 18 kg weight, but he has lifted a range of lighter and heavier weights, he may be rationally confident that he can lift the 18 kg weight, given his observation that his abilities generally don't have weird gaps. Witnessing the success and failure of others may also help. If J sees that similar agents easily lift 20 kg weights, he may infer that he can do the same – that he, too, would lift the weight if he chose the right motor command.

More specialized sources of evidence may also play a role. Massin (this volume) points out that our experience of the effort required to perform an act helps us gauge our general capacities. If lifting

20 kg feels easy, you have reason to think that you could lift a 25 kg weight. Similarly, Spener (this volume) suggests that "metacognitive" feelings of confidence help gauge one's capacities. If you feel confident that you are solving a calculus problem correctly, you have reason to think that you can solve even harder problems. I am happy to grant these points. But I would insist that the relevant feelings do not provide a self-standing epistemic basis for the relevant knowledge. It is a contingent empirical fact that if lifting 20 kg feels easy then lifting 25 kg (or even 20.1 kg) is possible.

The story I have sketched is empiricist and *actualist*, in the sense of Maier (this volume): Our knowledge of what we can do is not based on direct insight into unexercised abilities, but on mundane evidence of occurrent events – most importantly, of past successes and failures.

But my story also has a non-empiricist and possibilist element. I have focused on how J may know that he would succeed if he tried to perform an act, by issuing the relevant motor command. But how does he know that he can try? How does J know that he can realize $X2$? This question, I have argued, may not have an answer. In the cognitive architecture I have sketched, the agent may simply lack the capacity to be unsure about the available motor commands. Knowledge of the available motor commands comes for free.

References

Arif Ahmed [2013]: "Causal Decision Theory: A Counterexample". *The Philosophical Review*, 122(2): 289–306

Maria Alvarez [2013]: "VI-Agency and Two-Way Powers". *Proceedings of the Aristotelian Society (Hardback)*, 113(1pt1): 101–121

Randolph Clarke, Justin Capes and Philip Swenson [2021]: "Incompatibilist (Nondeterministic) Theories of Free Will". In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy,* Metaphysics Research Lab, Stanford University, fall 2021 edition

Cian Dorr [2016]: "Against Counterfactual Miracles". *Philosophical Review*, 125(2): 241–286

Peter A. Graham [2011]: "'Ought'and Ability". *Philosophical Review*, 120(3): 337–382

Brian Hedden [2012]: "Options and the Subjective Ought". *Philosophical Studies*, 158(2): 343–360

Jenann Ismael [2007]: "Freedom, Compulsion and Causation". *Psyche*, 13(1): 1–11

Frank Jackson [1991]: "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection". *Ethics*, 101(3): 461–482

James Joyce [1999]: *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press

Immanuel Kant [1781]: *Critique of Pure Reason*

Angelika Kratzer [1977]: "What 'must' and 'can' Must and Can Mean". *Linguistics and Philosophy*, 1(3): 337–355

Keith Lehrer [1968]: “Cans without Ifs”. *Analysis*, 29(1): 29–32

David Lewis [1979]: “Counterfactual Dependence and Time’s Arrow”. *Noûs*, 13: 455–476

— [1996]: “Elusive Knowledge”. *Australasian Journal of Philosophy*, 74: 549–567

John Maier [2018]: “Abilities”. In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy,* Metaphysics Research Lab, Stanford University, spring 2018 edition

H.A. Prichard [1949]: “Duty and Ignorance of Fact”. In W.D. Ross (Ed.) *Moral Obligation,* Oxford: Clarendon Press, 84–101

Frank Plumpton Ramsey [1931]: “General Propositions and Causality”

W. David Ross [1939]: *Foundations of Ethics*. Oxford: Clarendon Press

Wolfgang Schwarz [2020]: “Ability and Possibility”. *Philosophers’ Imprint*, 20

— [2021]: “Objects of Choice”. *Mind*, 130(517): 165–197

Jordan Howard Sobel [1983]: “Expected Utilities and Rational Actions and Choices”. *Theoria*, 49: 159–183

Toby Charles Penhallurick Solomon [2021]: “Causal Decision Theory’s Predetermination Problem”. *Synthese*, 198(6): 5623–5654

Peter van Inwagen [1983]: *An Essay on Free Will*. Oxford: Clarendon Press

Jonathan Way and Daniel Whiting [2016]: “If You Justifiably Believe That You Ought to Φ, You Ought to Φ”. *Philosophical Studies*, 173(7): 1873–1895

Paul Weirich [1983]: “A Decision Maker’s Options”. *Philosophical Studies*, 44(2): 175–186

Michael Zimmerman [2008]: *Living with Uncertainty: The Moral Significance of Ignorance*. Cambridge University Press