International Theory (2015), 7:2, 231–262 © Cambridge University Press, 2015
doi:10.1017/S175297191500007X First published online 29 April 2015

Why did they do that?: the methodology of reasons for action

JOSEPH O'MAHONEY
School of Diplomacy and International Relations, Seton Hall University, South Orange, NJ, USA

E-mail: joseph.omahoney@shu.edu

'Why did they do that?' is one of the most common questions in International Relations. However, we cannot access other people's reasons for action the same way that we perceive our own; we cannot introspect the reasons of other actors. This paper provides a unifying framework that delineates different types of knowledge claims regarding reason attribution. There are three possible methodological responses: (1) assume a possible reason and explain behavior in terms of that reason; (2) avoid the direct attribution of reason to individuals and locate explanatory leverage at an analytical level beyond the individual actor reason; and (3) use empirical evidence to adjudicate between possible reasons. Excessive skepticism of evidence of reasons lessens our understanding of the causes of action. When using empirical evidence, contrary to existing arguments, the paper shows that private settings do not systematically favor the true revelation of reasons. The paper also proposes a general principle, consilience, that allows evaluation of empirical claims of reason attribution that subsumes several existing methodological considerations, organizes them, and gives a consistent means of choosing between alternative reason attributions.

Keywords: philosophy of science; epistemology; methodology; reasons for action; explanation

'Why did they do that?' is one of the most common questions that political scientists seek to answer. The idea of a reason behind an action is an important part of the debate surrounding causation in the social sciences and in International Relations (IR).[1] The word 'cause' originally meant a reason for an action, coming from the Latin word *causa* meaning 'purpose' or 'reason' (Martin 2011, 30). In practice, most explanations of behavior in political science rely, whether implicitly or explicitly, on some attribution of

[1] Not all claims of social science causation involve attributing reasons for action, as I show below.

motive, intention, or some reason for action to actors.[2] Explanations that do *not* give a reason why the actors involved did the action that we are trying to explain feel unsatisfying. However, there is a tension involved when we try to justify how we know what the reasons of actors are. We cannot see or perceive other people's reasons for action the same way that we see our own; we cannot introspect the reasons for action of other actors. This presents us with a methodological problem that, while well known, has no systematic treatment in the IR literature.[3]

Existing reactions to this methodological problem, the fundamental problem of reason attribution (FPORA), has had several negative effects on the current IR literature that this paper addresses. First, blanket skepticism about reasons, or the idea that we can never really know why people did what they did, is too extreme a reaction to the fact that evidence is imperfect.[4] Such excessive skepticism comes from across the theoretical spectrum, from rationalism (Frieden 1999) to constructivism (Krebs and Jackson 2007). At worst, this skepticism can lead to denying that we can ever know actors' true reasons and at the same time relying on reasons as a core part of your theoretical apparatus. This combination is exemplified by one version of the strategic choice approach to international relations (Frieden 1999), which holds that preferences and beliefs are essential to understanding, but that scholars should not try to observe them. Another problem is that skepticism about reasons may provoke unnecessary unambitious theorizing via trying to avoid relying upon reasons.

A second problem is a lack of general and consistent rules for evaluating competing reason attributions. Despite the centrality of this problem to explanation, and the fact that multifarious insights on ways to approach this problem are scattered around the current literature, there is no clear specification of the types of claims that political scientists and IR scholars want to make about reasons. Without a clear idea that there are multiple different enterprises that we can be engaged in, conflation of types of knowledge claims can lead to criticisms that do not actually apply, erroneously undermining reason attribution. This can lead to uneven and

[2] Most of the interest in IR is in attributing reasons to historical actors on the basis of documentary evidence. Therefore, I focus on that situation in this paper. I also deal with reasons for action of individual people, rather than corporate actors. In practice, discussion in IR often treats states as individual actors, but this issue is beyond the scope of this paper.

[3] For example, in a canonical methods text, King *et al.* (1994, 110) mention this problem but other than recommending specifying observable implications do not provide detailed guidance on how to address it.

[4] That we can never really know for certain is true of all empirical claims about anything and is not unique to reason attribution.

unconvincing analyses as well as fueling the belief that evidence of reasons is uniquely problematic. Clear separation between these types of knowledge claims could facilitate appropriate evaluation of the claims that scholars actually make.

Third, the misidentification of the contribution of purely theoretical explorations of reasons both undervalues theoretical work without an empirical component and oversells the empirical validity of, for example, formal modeling.

In this paper, I provide a unifying framework that delineates different types of knowledge claims regarding reason attribution. The paper articulates the FPORA and three possible methodological responses: (1) assume a possible reason and explain behavior in terms of that reason; (2) avoid the direct attribution of reason to individuals and locate explanatory leverage at an analytical level beyond the individual actor reason for action; and (3) use empirical evidence to adjudicate between possible reasons. I show that how scholars react to the fundamental problem has constraining effects on the sort of knowledge claims that they can consistently make. I argue that assuming or avoiding reasons because of excessive skepticism about empirical evidence of reasons means losing a valuable part of our understanding of the causes of action.

Existing mainstream methodological advice on reason attribution is largely limited to exhortations to use private documents rather than public statements, or occasionally to pay attention to the context of a speech act. However, the private documents gambit is not a panacea and may even be the *opposite* of good practice, depending on the type of knowledge claim. In addition, consideration of context is a largely *ad hoc* maneuver, often only specifiable in particular cases, and without any guidance on how to evaluate alternative claims regarding the role of context. This paper specifies the knowledge claims for which private documents are potentially preferable and argues, contrary to existing claims, that private settings do not systematically favor the true revelation of reasons. The paper also proposes a general principle, consilience, that allows evaluation of empirical claims of reason attribution and that subsumes several existing methodological considerations, organizes them, and gives a consistent means of choosing between alternative reason attributions.

This paper proceeds as follows. First, I address the definition of reasons for action and introduce the use of reasons in causal explanation. Then I lay out the FPORA. The paper then considers and critiques two potential responses to this problem; assuming reasons while still using them in your causal claims, and avoiding using reasons in causal claims. Finally, I address the use of empirical evidence in reason attribution, considering a variety of issues.

Reasons for action

One type of explanation in social science involves the attribution of a reason to an actor. The actor performed an action because of the reason for that action. Davidson (1963) makes the argument that the cause of an action can be described as the agent's reason for doing what she did. This he calls the primary reason for the action. A primary reason is made up of a 'pro-attitude', that is, what it was about the type of action that appealed to the agent, and a belief that the action was of that type. When we explain an action by providing the primary reason, we 'rationalize' the action, that is, make it understandable.[5] Reasons need not be rational. For example, Lebow contends that 'most, if not all, foreign-policy behavior can be reduced to three fundamental motives: fear, interest, and honor' (2010, 14). Taylor similarly prioritizes understanding the reason an agent has for action:

> Now insofar as we are talking about behavior as action, hence in terms of meaning, the category of sense or coherence must apply to it. This is not to say that all behavior must 'make sense', if we mean by this be rational, avoid contradiction, confusion of purpose, and the like. Plainly a great deal of our action falls short of this goal. But in another sense, even contradictory, irrational action is 'made sense of,' when we understand why it was engaged in. We make sense of action when there is a coherence between the actions of the agent and the meaning of his situation for him. We find his action puzzling until we find such a coherence. It may not be bad to repeat that this coherence in no way implies that the action is rational: the meaning of a situation for an agent may be full of confusion and contradiction; but the adequate depiction of this contradiction makes sense of it (Taylor 1971, 13–14).

Making an action comprehensible, or 'make sense', involves redescribing the reason for the action in a way that intuitively makes sense to the observer. For example, I can explain Jim's going to the store by saying that Jim was thirsty and wanted to buy a drink. This makes the action comprehensible. If instead, I said that Jim went to the store because he wanted to see what the moon was like, I have not made the action comprehensible.[6]

[5] Davidson's position has attained quasi-consensus status in philosophy, although there is still contestation over philosophical issues, such as Davidson's anomalous monism, whether mental causation is compatible with free will, whether causation is a single thing or an umbrella term collecting multiple types of relations, and whether types of mental states are identical with types of physical states (see D'Oro and Sandis 2013).

[6] This account of comprehensibility is merely illustrative of the central point. There are other issues involved in comprehensibility such as the requirement of an intersubjectively shared meaning of terms, and the background knowledge necessary for such a shared meaning. These issues are not the focus of this paper.

In this paper, the term 'reason for action' or 'reason' is used to refer to intentional mental states and their components.[7] There are a variety of terms that are sometimes used interchangeably in IR to refer to the contents of people's minds before and during action. One common distinction is between preferences (or interests or desires or goals) and beliefs, which is similar to Davidson's distinction between pro-attitudes and beliefs.[8] Other terms used include motivation and intention. To say that an actor is motivated to do something incorporates the idea that the action is goal directed, but it is possible to have a motive but not act upon it.[9] Bratman makes the case that desires and beliefs are insufficient for action and that intentions are also required. Intentions are 'distinct psychological elements' (Bratman 1981, 263) that essentially involve the formation of plans of action connecting desires and beliefs together in specific ways. In IR, Rosato (2014, 53) argues that state intentions are analogous to 'strategies' in game-theoretic terminology, although without the technical baggage. In order to encompass this variety of mental states, I use the broad term 'reason for action'.

For ease of expression, in this paper, I treat the reason for an action as a single entity. However, multiple reasons are not only possible but plausibly endemic (Jansz 1996, 480). Reasons can be multiple in two ways: nested and overlapping. Nested reasons are reasons at different levels of abstraction where each more concrete reason is an instance of a more abstract reason. Baldwin illustrates this idea while discussing the reasons for the action of enacting a tariff on autos, with six levels of abstraction:

1. Getting Japan to export fewer cars to the United States, which is in turn a means to
2. Supporting the price of domestically made autos, which is in turn a means to
3. Ensuring the survival of the domestic automobile industry, which is in turn a means to
4. Promoting the US 'national interest', which is in turn a means to

[7] That is, intentional in the sense of intending to do something, not intentional in the sense of representing or directedness. An example of a non-intentional mental state might be aimless day dreaming.

[8] Davidson lists, 'desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values' as examples of pro-attitudes (1963, 686).

[9] In ordinary language the word motive is often used to refer to an ulterior motive, that is, one different from the one that is apparently driving someone's behavior (as Berard 1998, 196–98 notes).

5. Serving God's will by saving the world from the scourge of atheistic communism, which is in turn a means to
6. Ensuring peace for one's soul in the hereafter (1985, 48).[10]

Overlapping reasons are reasons whose object is the same. For example, invading another country might be attractive because you think it demonstrates your capability and resolve, satisfies domestic desire for action, and increases the resources you have direct access to.[11]

There are challenges to the use of reasons in explanation. Apart from the methodological challenges discussed in this paper, some objections reject the causal role of motives. A systematic evaluation is beyond the scope of this paper, but I mention several such arguments here. One is that conscious motives do not cause behavior because they do not really exist, another is that subconscious psychological processes do so instead.[12] Another objection is that mental states of individuals add nothing to explanation because the meaning of action is determined intersubjectively and not by an isolated individual. The thrust behind this argument is that what an action is or counts as, that is, the meaning of an action, is often not simply dependent upon the intentions of the actor. For example, I can insult you without intending to do so. Explaining what happened in this situation would involve reference to wider social structures and rules, like what counts as an insult in the society, what the appropriate response to an insult is, and so on. Knowing my mental state immediately before my action does not, in and of itself, determine the description of the action as an insult, nor does it fully explain why this physical action or speech act can provoke such an extreme response.[13] However, this objection fails to appreciate that explaining why someone did what they did is more fine-grained than merely accurately describing the action.[14]

[10] A methodological problem nested reasons pose is that the theory choice at issue might be specified at a lower level of abstraction than the evidence. The evidence might then not be able to choose between alternatives.

[11] One challenge for reason attribution is that the actual reason may be one, two, or all of these together, in that the action would not have been performed unless multiple reasons all pointed in the same direction. This problem can be mitigated by treating the set of overlapping reasons as a single-reason complex.

[12] For example, Lichtenstein and Slovic (2006) on the idea that preferences are constructed in the process of elicitation, and Butler (1990) on the idea that performing certain actions construct a sense of psychological interiority or subjectivity, which is fictional.

[13] Thanks to an anonymous reviewer for raising this point (see also Moon 1975, 161–71).

[14] With the insult example, there are multiple possibilities. One is that I try to insult you and then you do not get insulted, so 'I insult you' is false. Another is that I say something unknowingly insulting and so 'I insult you' is true, even though my mental state was not 'intend to insult'. Another is that I say something that is an insult according to the linguistic and cultural institutions of our society, but I do not mean it as an insult nor do you take it as such (e.g. a joke). In all these

Another variety of reason skepticism resorts to the idea that it is irrelevant whether we make correct reason attributions, as long as our predictions about behavior are correct. This is known as the 'as if' approach (Friedman 1953). However, quite apart from the fact that the empirical performance of the 'as if' theories is lacking (this was part of the impetus behind the field of behavioral economics, for example) the criterion of success involves prediction and not explanation.

Regardless of such challenges, where it is often not made explicit, much social science explanation is premised on an acceptance of the Davidsonian premise that providing a reason for action explains the action. Prominently, rational choice analysis requires that an actor made a choice for a reason and this reason explains why the choice was made. For example, Frieden, outlining the strategic choice approach to IR, argues that specification of preferences is essential to a 'full understanding of the sources of international political outcomes' (1999, 53). Aside from rational choice, many approaches are based on the idea that reasons are a primary building block of explanation. Elster (2007) only accepts explanations framed in terms of intentional choice. Intentionalist interpretivism also constructs explanations in 'terms of intentional explanation, i.e. in terms of the outcome (or aggregated outcome) of identifiable political actors acting upon intentional states they might plausibly hold' (Adcock 2003). Weber (1978, 11) placed understanding the motives of actors, defined as complexes of meaning that seem to the actors an adequate ground for the conduct in question, at the center of a scientific approach to social explanation. Aggregating such reasons for action can provide explanations for even highly complex social phenomena (Hayek 1980; Schelling 1978).

The FPORA

Davidson's argument brackets any methodological issues, but he does make a distinction between a rationalization that provides the *actual* reason why the action was performed and one that is reasonable/plausible, but does not appeal to the reason that was the agent's actual reason. This distinction is crucial for any methodological discussion of motives. From this distinction arises an issue that is common to all social sciences, which I call the fundamental problem of reason attribution:

FPORA: It is impossible to directly perceive what someone else is thinking.

This is also true in daily life. Why my partner, colleagues, parents, or a stranger acts the way they do can only be attributed on the basis of

situations, understanding the causes of the actions requires reference to the subjective intentions of the actors.

https://doi.org/10.1017/S175297191500007X Published online by Cambridge University Press

observation of their behavior, including speech acts. I think I know why I am doing something. By using introspection, or self-examination of our mental states, we have a mode of access to our own mental states that is unavailable when we are trying to figure out the motives of other people. I can introspect my reason for leaving a conversation with a colleague early, whether it be that I do not want to be late for a meeting or that I am uninterested in their conversation. When that same colleague stops our conversation short, I am left to deal with the FPORA.

There is a history in IR of a widely perceived problem stemming from the roots of the discipline in a debate between idealism and realism, encapsulated in the well-known difference between the logic of consequences and the logic of appropriateness (March and Olsen 1998). Numerous middle-range debates, such as the one about greed vs. grievance in civil war studies, or whether the power of the UN Security Council is about information revelation or normative legitimacy, are fundamentally about reason attribution. This is the understandable doubt that a political actor 'really believes' what he/she is saying when they appeal to altruistic or moral values. For example, Lebovic and Voeten (2006) claim that 'the empirical implications' of an argument that states shame human rights violators because it provides information on a state's reputation for good and bad behavior and an argument that such shaming is done because non-violators morally disapprove of violators are 'difficult to disentangle'. This doubt elides the distinction between different rationalist reasons and reifies the distinction between two classes of reason. The more fundamental issue is the FPORA.

Introspection is not foolproof. Humans can suffer from self-deception, or 'denial' of their true motivations. This can be owing to the desire to conform our thoughts to socially acceptable reasons so as to preserve our sense of self-worth. Some research suggests that we lack the vocabulary to describe some of our thoughts, feelings, or attitudes and so we borrow labels and models from society to help us understand our own mental processes (Wittgenstein 1953; D'Andrade and Strauss 1992; Jansz 1996). A line of psychological research investigating self-perception theory has found repeatedly that a behavior comes *before* a related emotion, despite people's self-reports based on introspection. If people are induced to smile they report being happier, if they are told to stare into a stranger's eyes they report more romantic attraction to the stranger, if they are told to stand up straight they report higher self-esteem than if told to stand slumped, and so on (Laird 2007). Another serious issue is confabulation, such as when people manufacture reasons for their behavior, even when they did not perform the behavior in the first place (Hirstein 2005). The conclusion reached is that the individuals themselves make mistakes when

introspecting the causes of their actions. These lines of research suggest that the FPORA can apply even to first-person reason attribution.[15]

How can we deal with the FPORA? In this paper, I distinguish between three types of approach to the FPORA: assumption, avoidance, and using empirical evidence. I lay out precisely the form of explanation underlying each type of approach and identify the attractions and limitations of each. Both the assumption and avoidance approaches have costs and rest upon over-stated objections to the use of evidence in reason attribution.

Methodological responses to the fundamental problem: assumption

One response to the FPORA is to assume or postulate a possible motive and only accept as valid explanations those that ultimately rely on that motive. This we can call the assumption response to the FPORA. This is a popular response in the IR and political science literature (not to mention economics). An empirical phenomenon, a behavior or a set of behaviors, are observed and then found puzzling. The analyst wonders how it came about. They then posit a reason that makes this action or set of actions comprehensible. For example, Fearon aims to explain why wars are costly, but, nonetheless, recur by characterizing the 'full set of rationalist explanations that are both theoretically coherent and empirically plausible' (1995, 380). Bueno de Mesquita *et al.* (2003) built a theory of domestic political institutions and foreign policy on the simple assumption that state leaders are motivated by a desire to survive or remain in office. Frieden makes a further distinction between assuming and 'deducing' preferences, that is, assuming preferences on the basis of intuitive plausibility alone, or having a 'prior theory of preferences' (1999, 61) that links actor preferences to actor properties or environment. For example, if we assume that firms are motivated by profit maximization and that their profits depend on some specification of how trade policy affects a type of firm's profits, we can 'derive' that type of firm's preferences over trade protection. However, this is not epistemologically separate from the assumption solution to the FPORA. Any confidence we should have in a knowledge claim about firm preferences over trade is a direct function of our confidence in the plausibility of the initially assumed preferences.

However, the implications of the assumption response may be more uncomfortable than they seem and conflict with current practice.

[15] Thanks to an anonymous reviewer for this suggestion.

What is the form of a knowledge claim under the assumption response?

A1: I am willing, by assumption, to accept as an explanation for actions of type A a reason *s* from the set of reasons S.

A2: Reason *s* in S renders action *a* of type A comprehensible to me.

A3: Therefore, *s* is an explanation for *a*.

Note that A3 does not say that *s* is *the* explanation for *a*. There are two consequences of this type of knowledge claim. First, unless the set S has only one element, then even knowledge claims under the assumption response require some method of distinguishing between different reasons. It is possible that two reasons from S make the action comprehensible. For example, a leader may sign a ceasefire because she wants to pursue peace and sees a ceasefire as a first step toward a lasting peace agreement, or she may want to reattack later to improve her sides' bargaining position and sees a ceasefire as buying her troops some time to rest and redeploy. One means of distinguishing between reasons is sheer plausibility. It may be the case that reason *s* is comprehensible to you, but you find it implausible in a particular circumstance. If this is a crucial part of the warrant for your knowledge claim, this should be explicit.

The plausibility of reasons for action

In one example of an explicit appeal to plausibility, Katznelson and Weingast (2005), commenting on the tendency of rational choice institutionalists studying the US Congress to assume the motive of maximizing the probability of reelection, defend the assumption as justified by its general plausibility. They reference Mayhew (1974), who concludes that the individualistic, open US electoral system, and a minimally restrictive party affiliation procedure provide an incentive to focus on reelection at the expense of, say, party loyalty or bureaucratic wrangling, as it might be in the United Kingdom or continental Europe.[16] However, Mayhew makes clear that this is contingent on a desire to make a career in Washington, something that Mayhew asserts was unusual in the pre-Civil War era. Even in Mayhew's study, there is no empirical evidence provided that any action by any individual was actually motivated by the desire for reelection, except a single generalized out-of-context quotation. Katznelson and Weingast's appeal to Mayhew's work thus does not mitigate the fact that their response to the FPORA is to assume the motivations of actors.

Under the assumption response, there are no grounds for resolving a challenge based on the superior plausibility of an alternative reason, *except* in terms

[16] This claim also rests on the premise that if you have a contextual incentive to want to pursue something, you are or will be motivated to act on that incentive. For a discussion, see below.

of plausibility. If any empirical evidence is used to decide between reasons, then the knowledge claim is subject to all of the issues inherent in such an enterprise (see below). This is a crucial point. Assessing plausibility, however it is done, is a separate task from weighing evidence in a reason attribution.

What does plausibility mean? Is it an unjustifiable intuition? Are there general properties of plausibility with regard to reason attribution, such as logical consistency or phenomenological familiarity? The key feature of rational choice explanations is that acceptable reasons must conform to some minimal standard of rationality, such as completeness and transitivity of preferences. Maybe there are some features, apart from any evidence about the particular case, of the current state of knowledge that mean that there is some sort of distance from our prior beliefs to the postulated reason. This is a question for future research.

The second consequence of the assumption response to the FPORA is that this form of warrant does *not* allow for any claims that the action is evidence supporting the truth of the reason *s* or the set of reasons S. Instead, the assumed reason is explaining the action to you. That is, the assumed reason makes the action comprehensible.

Imagine that I see the United States limiting its power by the construction of international organizations (IOs) that hamper its freedom of action (Ikenberry 2001). This is puzzling because it seems as if freedom and flexibility are things that states should pursue. Ikenberry dissolves this puzzlement and explains the action by saying that the United States was trying to transmit credible information of benign intentions to other states in the system. If I am resorting to the assumption response to the FPORA, then this is as far as I can go. It would be inconsistent to use statements of intention to communicate benign intentions as supporting evidence that was the motivation because then I would be appealing to a different warrant for my knowledge claim, that is, using evidence to adjudicate between possible reason attributions. However, under the assumption response, it *is* consistent to claim that statements of other reasons are irrelevant to your knowledge claim. That is, if you are relying on the assumption response, then statements by participants that they were constructing IOs in order to, say, exert control in particular areas of international activity, or to placate a domestic constituency, do not affect your knowledge claim. Rather than using evidence to decide between different possibilities, explanations based on the assumption response explain actions in terms of the assumed reasons. All of the analytic work happens at the level of the logical consistency or the plausibility of the attributed reason.

The costs and benefits of the assumption response

Why would you want to merely assume reasons? A possible objection is that to only assume or postulate reasons is not a valuable task. One major

limitation of simply assuming a reason is that there is no claim that the 'explanatory' reason is the *actual* reason driving behavior. It thus seems to be merely a preliminary step before the valuable part of research, the empirical investigation. Clarke and Primo (2012) argue against this line of thinking. They point out that elucidating a model, such as a coherent set of reasons for action, constitutes a valuable use of time and effort. Rubinstein (1991) agrees, pointing out that the value of game theory is that it aids in the exploration of reasoning in strategic settings. Schelling reports a reader who had 'simply not comprehended that an inherently non-zero-sum conflict could exist' before reading theoretical elaboration of such a scenario in Schelling's (1980, vi) work. Laying out or elaborating a reason or set of reasons and its implications clearly, explicitly, and in detail, is hard and uses different skills from those employed in empirical work. Elaboration of reasons for action may also suggest sources of variation in action that would not have been apparent otherwise.

In addition, the evidence available for or against a particular reason may be non-existent or so sparse and underdetermining that simply explaining an action may be the best we can do. However, there can be serious costs to not engaging with the available evidence. Wendt points out that knowing how actors 'were actually thinking and motivated' is central to choosing between equally plausible alternative explanations for international institutional design choices. That is, in order to get to 'the real explanation … we need to get inside the heads and discourse of decision makers and see what is motivating their behavior' (Wendt 2001, 1028). An excellent example of the drawbacks of merely assuming reasons is the audience costs literature. Since Fearon's (1994) elaboration, there has been a profusion of work utilizing the idea that a crucial way state leaders can make credible threats is to engage domestic political costs, 'audience costs', for backing down if the threat is not followed through on. This idea solved several theoretical and empirical puzzles, including providing a mechanism for the democratic peace finding (Schultz 2012). However, recent work has empirically investigated whether the audience costs mechanism actually motivates crisis behavior. Trachtenberg (2012) studies whether state leaders intentionally try to use audience costs to lock in a bargaining position and whether the adversary understands that the threatening state leader 'would find it hard to give way for fear of incurring audience costs'. His question is do governments 'actually make this kind of calculation?' (Trachtenberg 2012, 7). Despite the plausibility of the reasons involved, Trachtenberg finds that audience costs do not motivate behavior in the cases he studies. Similarly, Snyder and Borghard (2011, 437) find 'hardly any' evidence that audience costs motivate behavior in post-1945 crises. Mercer encapsulates the primary drawback of the assumption solution: 'Although one might use audience costs to solve a variety of

puzzles, if audience cost mechanisms are imaginary, then so are the solutions that rely on them' (2012, 399).

Avoiding direct reason attribution

Another response to the FPORA is to avoid relying on reasons as part of your explanatory strategy. I call this the avoidance response. The assumption approach is still fundamentally engaged in the enterprise of reason attribution. The avoidance approach, however, is not. Instead, recourse to the avoidance approach entails explicitly avoiding *any* reliance on reasons as explanatory. This means that, as long as the explanatory goal is consistent with the claim, strictly speaking reasons are irrelevant to explanations taking the avoidance approach. One key benefit of the avoidance approach is that the uncertainty surrounding how we can know people's reasons is no longer a problem, that is, the FPORA does not apply. However, there are costs to setting our explanatory sights lower than they often need be.

Defining the avoidance approach

The perceived need for intentional explanation is strong; however, there are other types of causes that are fruitfully hunted in political science. Many causal inferences have been made without relying on an answer to the question of reasons for action. One example is Fearon and Laitin's (2003) work on the duration of civil wars. Among other claims, they argue that mountainous terrain, a proxy for the availability of hard-to-eliminate rebel safe havens, causes an increase in the duration of civil wars. We can use this to explain why some civil wars last so long, or longer when compared with others. This argument is compatible with numerous reasons for pursuing rebellion or for government counter-insurgency actions and thus does not rely on any one particular reason. Similarly, some explanations rely upon network analysis to explain patterns of behavior. Nexon explains the seemingly radical shift in allegiances from the Taliban to the United States in Afghanistan after 2001 with reference to dependency in patron–client relationships; because warlords were bribed to switch, most of their followers switched with them. Clients may have acted out of loyalty, honor, or calculation of the benefits of not breaking ties with the patron. This claim about social ties 'provides significant explanatory power independent of specific microfoundations' (Nexon 2009, 41). The key underlying characteristic of the avoidance response is that the knowledge claim must be fully compatible with more than one (themselves incompatible) reason attributions:

> An explanation ε of action a avoids the attribution of reasons if there are two (or more) reasons such that s can be true when ε is true, and s^* can be true when ε is true, but s and s^* cannot both be true at the same time.

https://doi.org/10.1017/S175297191500007X Published online by Cambridge University Press

An innovative recent line of work that relies upon the avoidance response to the FPORA appeals to the reasons and justifications for action that individuals express, but avoids the attribution of a specific reason to an individual at a particular time. Instead, the constellation of explicit justifications for an action is charted, with the analyst coming to a decision on which justifications were *socially sustainable* (see also Mills 1940). Jackson (2006, 42) argues that the collective shape (or topography) of the legitimation debate over a policy or set of policies is causally sufficient for an outcome. Not only are policy makers restricted to enacting 'those policies that they can justify in a manner acceptable to their audience', but the ways in which a policy is justified 'makes the policy proposal possible and helps it to win out over alternative courses of action' (Jackson 2006, 25, 29). This position is premised on agnosticism as to the reasons inside individual actor's heads, and instead operationalizes the acceptability of policies in terms of how other actors speak and write about, or act toward the policy. The argument is formalized and extended in the study by Krebs and Jackson (2007). In an explicit rejection of reasons as a response to the FPORA, Krebs and Jackson (2007, 41) allow that actors possess motives that shape behavior relevant to political outcomes; however, they propose a type of explanation to which 'purposive accounts' (i.e. reason attributions) are irrelevant.

Reasons here are still explanatory, but not in the sense that actor X did action *a* at time *t* for reason *s*. Instead of this sort of judgment, a judgment is made as to what the articulated justifications for action were and which of these were *accepted* by the relevant actors in the sense of not resisted. Regardless of what the individual motivations actually were, or whether reasons are constituted by internal mental states, this form of the avoidance response allows the analyst to use evidence to determine what the actors involved thought a convincing justification would be. Given that the justification was provided, it must have been thought convincing to some actor, even if that is the same actor who is expressing the justification. An actor makes a justification for an action or provides a reason why an action should be done in order to legitimate that action in the eyes of others. The aggregation of those reasons at a historical point in time constitutes the boundaries of legitimate discourse at that time for that action. Identifying socially sustainable justifications is especially useful when a group of actors are trying to come to a decision on a joint action, like in a government or a group of governmental actors. Individuals are being honest about their reasoning, or they are trying to be convincing to others, or they are trying to avoid censure for violating shared rhetorical standards. In all three of these situations, speech acts are potential evidence of the reasons that are thought convincing.

The form of the knowledge claim for a socially sustainable justification is:

SS1: A justification *j* for action *a* is given or appealed to by an actor X

SS2: *j* is either not challenged/disagreed with, or it is ultimately acquiesced to by actors y_i

SS3: Therefore, *a* was allowed to occur because of *j*.

Notice that individual reasons for action do not appear in this knowledge claim. The actors could be motivated by *j*, they could simply be willing to accept *j* as a reason for *a*, or they could be motivated by *j**, but know that others would not be willing to accept *j** as a reason for action.

Here the distinction between private and public justifications is important. As public pronouncements are the only ones that are being used to legitimate action, they are more important than private statements as evidence to a socially sustainable justification explanation.

The costs of avoidance

Is anything lost by recourse to an avoidance explanation? That is, can investigating reasons in an avoidance explanation add anything to our understanding of the causes of action? Let us return to Fearon and Laitin's civil war example, in which we know that mountainous terrain increases length of civil wars on average. It would be a distinct improvement in our causal understanding to also know that the rebels retreating to mountain hideouts do so because they intend to regather their strength for future fighting, rather than that they want to surrender, but they have to hide from authorities intent on eradicating them. Similarly, Krebs and Jackson's claim that conscription helped the Druze get awarded Israeli citizenship would be supplemented by knowing what the Israeli authorities' motives were. Did they hope that including Druze in the Israeli Defense Force would bolster their claim to liberalism and undercut international criticism (Krebs and Jackson 2007, 50)? Or something else? Advocates of analyzing international practices have appealed to the 'great advantage of ridding [the analyst] of the need to make problematic claims about the state of mind amongst the people who perform the practice' (Andersen and Neumann 2012, 458). As Ringmar explains, however, merely describing the practices gives us 'no clue' as to the intentions behind the practices. Implicitly conceding this point, some practices analysts deal with this problem by adding in a reconstruction of the intentions and aims behind practices, for example, of deterrence in the Cold War (Ringmar 2014, 13). The point here is not that explanation is impossible without reasons. It is that reasons add to our understanding of the causes of action.

There are at least three ways in which avoiding reason attribution lessens our understanding of the causes of action. First, obviously, even though we may have a coherent and empirically supported explanation of an outcome, we do not know why the individuals involved performed the actions making up the outcome. However, there are other consequences important even if we are locating explanatory leverage elsewhere. The second cost of avoidance is the loss of some sources of variation in the actions of interest. For example, rebels regathering their strength might be less amenable to a negotiated settlement than rebels looking to surrender. An Israeli government on the lookout for legitimation in terms of international liberalism may do different things from a government looking to recruit the fiercest fighters. Third, avoidance also means losing a way to generate new theoretical arguments. Empirical investigation of a particular type of action might unearth a hitherto unimagined reason or perhaps one that is well known, but that no one had previously thought was driving this type of action. For example, the actions of US President Richard Nixon and his adviser Henry Kissinger during the 1971 war between India and Pakistan can only be explained with reference to their secret diplomacy with the People's Republic of China, ultimately motivated by grand strategy with regard to the USSR. To others, this motivation was unthinkable during the crisis surrounding the war (Hollen 1980).

The use of empirical evidence

The third type of response to the FPORA is to adjudicate between possible reasons for action with reference to empirical evidence. Despite the excessive skepticism of some in the political science and IR literature, humans are surprisingly successful at determining the mental states of others. There are whole literatures in philosophy, psychology, and cognitive science dedicated to exploring the fact that we, humans, cannot observe others' mental states, but, nevertheless, routinely use them successfully to explain and predict behavior. This capacity to judge others' intentions becomes apparent in young children and adults 'display consummate skill' at this task, perhaps suggesting 'neural structures innately equipped … to detect intentionality' (Baldwin and Baird 2001, 172, 176). There is much debate over whether we are able to directly perceive others' minds. Some hold that we can directly perceive mentality in others, that our phenomenological experience of others' mental states is immediate and not via the interpretation of the bodily movements of another person (e.g. Zahavi and Gallagher 2008). Another idea is that some aspects of extended or distributed cognition, like the use of gestures and or pen and paper for informational offloading, can be seen directly (Krueger 2012, 157). The alternative is to hold that

humans either have some sort of folk-psychological theory about how other humans think (theory theory) or that we build an internal simulation of others' minds (simulation theory) and use that to judge others' mental states. Bohl and Gangopadhyay point out that it is possible to doubt that we perceive shame, but not to doubt that we see someone's face being red, which seems to suggest that we do not directly perceive mental states. However, it also seems wrong to say that, for example, seeing a tree branch blowing in the wind and seeing someone wave goodbye are no different in terms of psychological experience (Bohl and Gangopadhyay 2014, 217). They conclude that other minds are not entirely hidden, but that seeing them is different in some sense from seeing the visual properties of other objects like color or shape.

Interpersonally, we have many cues of body language, tone of voice, maybe our previous experience of the person, that we can use to supplement any analysis of speech or action in a reason attribution. Schilbach *et al.* (2013), whereas proposing a new way of approaching the neuroscience of 'social cognition', argue both that the more emotionally engaged we are with another person's states or actions and the more we are in interaction with that other person, the more likely we are to be successful at knowing what that person was feeling or thinking. However, even in person we can be either unsure of reasons or even flat out wrong. St Thérèse of Lisieux, a Catholic saint, kept a journal of her time in a convent, in which she mentions another nun who 'annoyed [Thérèse] in all that she did'. Despite this antipathy, Thérèse was unusually nice to her colleague as part of her understanding of what was pleasing to God. The colleague, despite interacting with Thérèse every day, was completely unaware of the motive behind Thérèse's behavior, one time saying 'with a beaming face: "My dear Soeur Thérèse, tell me what attraction you find in me, for whenever we meet, you greet me with such a sweet smile"' (Lisieux 2005, XI, 31). I discuss the potential problem of misrepresentation in more detail later in this section.

The form of an empirical response to the FPORA is:

E1: *s* is a possible reason for actions of type A

E2: Given some evidence about what people said and/or did, I judge the most warranted reason for action *a* of type A at time *t* by actor X to be *s*.

Notice that in this formulation the action is a particular action, not a type of action. If you hold that the reason–action connection is characterized by equifinality [where multiple different causes produce the same type of outcome (Mahoney 2008, 424)], then establishing reasons in particular cases is logically before establishing whether *all* actions of type A are motivated by *s*, or whether they are more likely to be motivated by *s* than other reasons.

If we cannot introspect the reasons of others, what kinds of evidence can we use in a reason attribution? Morgenthau famously said that 'motives are the most elusive of psychological data' (1993, 5). In the following sections, I consider some of the issues involved in using three types of evidence: (1) statements, (2) actions, and (3) strategic context when supporting a reason attribution. I will be appealing to ideas from the discussion of *Inference to the Best Explanation* in the philosophy of science.[17] Inference to the best explanation, also called abduction, is a type of inference that privileges explanatory considerations, as contrasted to, for example, predictive success or logical deduction, in choosing between alternative theories (Lipton 1991). For example, inference to the best explanation appeals to 'theoretical virtues' as relevant considerations in theory choice. One of these is consilience,[18] 'the capacity to explain diverse independent classes of facts', or the idea that a theory or hypothesis gains in credibility to the extent that the several pieces of evidence in its favor are unrelated, which is widely held to be a valuable theoretical virtue (McGrew 2003, 561). Generally, the quality of consilience can be a guide to empirical practice. How can we apply this principle to reason attribution? First, there is a difference between consilience within particular classes of evidence (such as those I have delineated here: statements, actions, and context), and between those classes. Therefore, within the class of statements, if a reason can explain, or is consistent with, a variety of types of statement, such as statements made within a government *and also* between governments, this reason is to be preferred to one that cannot. Further, if a reason is consistent with a context (i.e. the strategic context provides an incentive for the actor and the reason incorporates that incentive), action (the reason explains the action or multiple actions), and statements (statements made by the actor or others indicate that the reason was relevant to the decision to perform the action), then this reason is to be preferred to one that is not. This is consilience across classes of facts. This principle of consilience collects some of our intuitive reactions to evidence, subsumes much of the existing methodological advice on reason attribution, and is a helpful guiding principle when we are trying to decide between reason attributions.

Statements of motivation

There is an intuitive plausibility to the following procedure. I want to know why actor X did action *a*. Actor X says, 'I did action *a* because I wanted to

[17] Philosophy of science should not dogmatically drive scientific practice (Gunnell 2011), but what it can do is shed light on what we are doing and make us wonder whether it makes sense.

[18] Thanks to David Waldner for suggesting this idea.

achieve outcome *b*, or for reason *s*'. Assuming that the given reason *s* is comprehensible to me, *s* explains *a*. This initially seems to be a best case scenario for empirically determining reasons for action. However, there is a prominent criticism of this procedure; the potential for dishonesty or misrepresentation in proclaiming one's reasons. The misrepresentation principle is:

> M: Any historical speech act of the form 'I did action *a* for reason *s*', is inadmissible as evidence that the reason for *a* was *s* because the historical performer of the speech act may have been lying.

There are two positions on dishonesty or misrepresentation regarding reasons. The first is that political actors' incentives for misrepresentation are so pervasive that it should be the default starting point of any analysis. The other is that misrepresentation should not be assumed but should instead be based on evidence, for example, some inconsistency between various pieces of evidence. Scholars often make a case for pervasive, systematic strategic misrepresentation that has *prima facie* considerations for reason attribution. For example, Mackie (1998) notes the tendency to assume that 'all men are liars' in political economy or public choice models of voting. Elster places this position front and center of a discussion of reason attribution. He points out that 'there are many reasons why people might want to misrepresent their motivations and those of their opponents' (Elster 2007, 59). For example, societies have a normative hierarchy of motivations and people gain an advantage by appearing to be motivated by a 'better' motivation (this is often a significant part of legitimation contests). However, one inherent problem with this argument is that it denies that we can know people's real motives by appealing to people's real motives, thus flirting with circularity (Bruce and Wallis 1983, 63).

Excessive weight is often placed in IR on the possibility of misrepresentation. This can have pernicious effects on research practice. For example, excessive skepticism of statements leads Simmons and Danner to ignore relevant evidence of reasons and instead resort to making claims based on wildly insufficient data. They ask why the International Criminal Court (ICC) was set up and 'more importantly, why do states agree to join this institution?' (Simmons and Danner 2010, 225). Noting that 'Evidence on governments' motive for joining the ICC is hard to come by', they assert that private statements are inaccessible and public statements are too coarse-grained, that is, they are consistent with numerous analytically distinct reasons. That said, they provide a couple of statements that the ICC will reduce conflict, restore confidence in the country acceding, and add to the accountability of the country's leaders. They even say of one statement, 'One can infer from this statement that the speaker acknowledges that

domestic processes are often less than effective, and that the ICC can in these cases provide a more effective—because more credible—substitute' (Simmons and Danner 2010, 237). However, they then assert, in a statement of excessive skepticism, that these statements 'reveal very little about the way supporters expect the Court actually to operate'. They then say that they will look at actions; however, they do not. Instead, they look at coarse-grained properties of states, like Freedom House democracy scores or whether a country is involved in a civil war, and infer directly from those properties the reasons for ratification. This evidence is insufficient to distinguish between alternative reason attributions, and so does not add to our understanding of the causes of ICC ratification. This is one example where excessive skepticism of statements has led to unconvincing reason attribution.

Historians attempt to mitigate the problem of misrepresentation with reference to expressions of reason from sources they deem less likely to be used for a political purpose, primarily private documents including classified internal memoranda and personal letters. As Elster notes, the underlying principle behind the credibility of these private documents is that they are 'less likely to be motivated by a desire for misrepresentation' (2007, 61). However, positing that some types of source [such as private letters and diaries (George and Bennett 2005, 100, fn 17)] are systematically less likely to be subject to misrepresentation is no more plausible than the reverse. That is, there is no *a priori* systematic difference between types of sources in terms of whether they more or less truly indicate reasons. As Broockman advocates, 'both public and private statements should be viewed through a strategic lens' (2012, 105).

Private settings qua private settings do not systematically favor the true revelation of reasons. If there is a strategic incentive to conceal or misrepresent reasons in the most public setting, there is also a strategic incentive to misrepresent reasons in other less public settings (Krebs and Jackson 2007, 40). Jacobs (2014) argues that, in general, political actors 'have incentives to exaggerate the importance of "good policy" motives and broad social benefits' of their policy positions. However, in general, political actors also have incentives to frame their behavior positively to any audience. If they are speaking privately to a small stakeholder group, they have an incentive to highlight the narrow individual benefits accruing to that group. If they are speaking to racists they are more likely to appeal to racist reasons than they are to non-racists.[19] It is easy to construct simple scenarios in which decision makers would not lay out their actual

[19] Thanks to Andrew Bennett for inspiring this point.

motivations in a private letter to a friend or family member. For example, maybe the actual reasoning is thought to be too complicated, morally questionable, or too reliant on special knowledge for the letter recipient to understand. If they are writing a letter to their spouse or family member, they have an incentive to make themselves seem as favorable as possible to that person, maybe by emphasizing the private benefits that they, and hence their family, will accrue from an action or policy. Even letters or memos to colleagues could fall prey to common distortions, such as a desire not to appear biased, idealistic, or naïve. For example, if they are writing an intragovernmental memo, they have an incentive to conceal their idealistic motives in favor of a cynical national interest-oriented analysis, so as to advance their standing and reputation for calm competence in that organization.

Such speculations could be multiplied. However, what they demonstrate is that in the absence of evidence, there is no good *prima facie* reason to accord one type of expression of reason primacy over another. Instead, misrepresentation should be a judgment arrived at on the basis of the evaluation of evidence. There is the principle of *nemo gratis mendax*, that is, no one lies freely. Under this principle, expressions of reason should only be doubted if there is a concrete reason to do so in a particular circumstance. That is, the construction of a reason for dishonesty or misrepresentation must use some sort of evidence. Otherwise, expressions of reason must be taken at face value. This principle focusses attention on specific reasons for doubt, rather than a generalized assumption that anything said in public by a political actor is necessarily dishonest or disingenuous. There is even some systematic empirical evidence that political leaders' public speeches accurately convey their actual beliefs. Utilizing the fact that US President John F. Kennedy recorded private discussions with key advisers in the summer of 1962, Renshon (2009, 656) compares the operational code (a set of political beliefs) of Kennedy's public and private speech, finding a 'striking similarity' between the two.

However, when we do have evidence of misrepresentation, then we should incorporate that into our reason attribution. One notable example of evidence of dishonesty is that with which Nexon opens his book on the political impact of the Protestant Reformation. Holy Roman Emperor Charles V, while publicly justifying his war against the Protestant states Hesse and Saxony as that they were 'transgressors of the peace against the Duke of Brunswick and his territory', wrote to his sister that this 'pretext will not long disguise the fact that this is a matter of religion, yet it serves for the present to divide the renegades' (quoted in Nexon 2009, 1). Here the fact that we have this evidence makes preferring the attribution of an ulterior motive, in the face of public statements otherwise, warranted.

The general principle of consilience is again a useful guide here. The attribution of reason misrepresentation only makes sense when there is an inconsistency between different classes of evidence, like different classes of statements. A good example of the utility of the consilience principle is Broockman's study of the negotiating positions behind the passage of a 1964 healthcare law in the United States. Broockman, addressing the 'problem of preferences', that is, that actors might strategically misrepresent their preferences during policy contestation and bargaining, advocates 'considering how actors' expressed preferences vary across strategic contexts' (2012, 84). Specifically, Broockman found that a previous study had misattributed the preferences of business interest groups on the basis of a letter from the National Association of Manufacturers (NAM) and Congressman John Byrnes, ranking Republican on the House Ways and Means committee. The letter includes reasons why industry would support a different version of the healthcare law than was currently being proposed, which the previous work had said meant that the NAM and business interests supported passing a version of Medicare. Broockman argues that evidence of statements from different strategic contexts, including from before the Democratic landslide victory in the 1964 election, shows that in fact business interests 'were totally and unmistakably opposed to Medicare' (2012, 93) and were proposing an alternative in order to dilute some provisions in the bill. Further, Byrnes' writings to other Republicans and other statements indicate that Byrnes proposed an alternative to Medicare, whereas expecting it to be rejected so that Republicans could convincingly argue that they were not opposed to healthcare for the aged. Here, in accordance with the consilience principle, Broockman's attribution of reasons to Byrnes and industry groups is superior because it is consistent with different classes of evidence, that is, statements from different strategic contexts.

Consilience can also accommodate the use of different classes of evidence of the *absence* of a particular set of beliefs. For example, when discussing the failure of German officials to consider the future misuse of funds collected for the first public pension scheme in 1889, Jacobs (2014) makes the case that the attribution of this state of mind is supported by the fact that there is no mention of it across several deliberative venues.

As intuitively attractive as statements of the form 'I did action *a* for reason *s*' are, much historical evidence is not as starkly related to the reason for an action. This means that, in practice, using statements as evidence of reasons requires an interpretation of those statements. Further, the conceptual vocabulary used by social scientists is often not the same as that used by historical actors in their self-ascriptions of reasons. As Taylor notes, in a discussion of how our representations, for example, justifications for

action, may not map perfectly onto the distinct categories of action created by social scientists, '… the person concerned may not even possess the appropriate descriptive term. For instance, when I stand respectfully and defer to you, I may not have the word "deference" in my vocabulary … This understanding is not, or is only imperfectly, captured in our representations' (Taylor 1993, 51). Using a statement or utterance as evidence in support of a reason attribution, requires making an interpretive judgment about the fit of that utterance into a particular theoretical framework. This can be a serious challenge, especially when dealing with actors and contexts with which we may be phenomenologically unfamiliar (see Bevir 1999 for a thorough evaluation of this and related issues). Skinner argues that one tactic for improving the quality of interpretive judgments is to 'trace the relations between the given utterance and this wider linguistic context as a means of decoding the intentions of the given writer'. Analysis of the social context is 'the ultimate framework for helping to decide what conventionally recognizable meanings it might in principle have been possible for someone to have intended to communicate' (Skinner 2002, 87). Hopf summarizes this position by saying that, 'Evidence does not consist of the actor's words alone' and that 'there must always be an accompanying account of the relevant sociohistorical context' (2007, 61).

Another interpretation issue is that statements sometimes cannot be accurately understood out of the context of the other statements made by an actor.[20] A statement may appear to indicate a particular reason or intention or belief but when viewed amongst other statements made by the same person, the statement may indicate something else. One example involves Stanley K. Hornbeck, the Chief of the Far Eastern Affairs Division at the US State Department in 1931. Hornbeck was considering potential reactions to the Japanese invasion of Manchuria earlier that year (the Manchurian Crisis). In particular, Hornbeck was evaluating the proposed policy of declaring that the United States would not recognize any result of the invasion. This policy later came to be called the Stimson Doctrine, after the US Secretary of State. Hornbeck wrote a memorandum, in which he writes that the proposed policy of non-recognition would, 'show the powers "mean business". It would give the Pact of Paris "teeth". It would answer the charge that the League and the various governments are impotent' (Doenecke 1981, 85). Several secondary sources use this quote as evidence that Hornbeck believed that the non-recognition policy would be effective. However, Hornbeck places 'mean business' and 'teeth' in

[20] George and Bennett (2005, 99–105) provide a thoughtful discussion of this particular issue.

quotation marks. This, in addition to numerous other statements made by Hornbeck that non-recognition would be relatively useless, suggests that he is responding to criticism made in the press and elsewhere, and that the policy of non-recognition might deflate that criticism.

This issue of the fallibility of interpretation places a heavy burden on the analyst, as establishing the context of a statement can be time consuming and difficult, and there may not be much available evidence. However, the evidentiary value of statements is dependent upon such interpretation.

Actions as evidence of reasons

Fear that actors misrepresent their true reasons for action is widespread. A common reaction to this problem is to rely instead on action as an indicator of reasons, dismissing expressions as 'cheap talk'. Only actions that incur enough costs to discriminate between actors *really* motivated by reason *s* and thus willing to bear those costs in the pursuit of *s* can be used as evidence of reason. This is the underlying logic of the resort to 'costly signaling' models from game theory (e.g. Fearon 1997). Elster, asking 'do they put their money where their mouth is?', uses the example of the Bush administration and its professed reasons for invading Iraq. The key stated reason was to seize or destroy Saddam Hussein's weapons of mass destruction (WMD). Elster (2007, 63) argues that a useful piece of evidence relevant to whether this was the true motivation would be whether the Bush administration took steps to protect the invasion forces from WMD. Here, the key principle is not the costliness of the action *per se*. Rather, if an action would only make sense if performed by someone with a particular reason, then it is good evidence in favor of that reason being the true reason. However, this strategy seems prey to the same sort of problem as afflicts the use of statements. Any action seems liable to a redescription that makes more than one reason equally supported. For example, poor planning, arguably in evidence in other respects during the Iraq War, could have meant that even a sincere Bush administration did not make adequate preparation for US troops to face WMD.

Another way of using actions as evidence of reason involves assessing if actions, other than the action to be explained, by the same actor can plausibly be described as motivated by the same motivation.

> C: If there are more plausible alternative descriptions of reason s^* for actions a^* of type A or even actions b of another type B taken by actor X, then s is not the reason for a.

In the abstract this can sound confusing and counterintuitive. However, this format for warranting a knowledge claim is widespread in lay

https://doi.org/10.1017/S175297191500007X Published online by Cambridge University Press

discussions of foreign policy and is sometimes used by scholars. For example, Girard (2004) asks why President Bill Clinton decided to invade Haiti in 1994. Girard uses the actions of the Clinton administration as evidence that the pursuit of democracy and human rights in Haiti were not primary reasons for the 1994 invasion, despite Clinton's explicit appeal to the need to restore democracy as a justification for the invasion in an address to the nation in September 1994. Girard points to the absence of US action to restore democracy in other countries, US willingness to accept the fraudulent 1994 electoral victory in the Dominican Republic in exchange for Dominican participation in the embargo on Haiti, and Clinton's refusal to ask for congressional approval of the invasion as evidence that the desire to restore democracy was not actually the motivation behind Operation Restore Democracy.

Underlying the idea that actions can be used as evidence of motivation for other actions is the concept of reason consistency. That is, people who are motivated to perform an action in pursuit of some goal are also motivated to perform other actions in pursuit of that goal. I have two reactions to this consistency premise. The first is that it is reasonable to suppose that an actor could be driven at two different times by two different reasons (this is another place where equifinality is relevant). I might want to eat spaghetti with marinara sauce now because it tastes nice, but later not eat spaghetti because I worry about embarrassing myself in a sensitive social situation. The latter action is not evidence against my eating spaghetti because I like it. My second reaction is that it is intuitive to hold some consistency requirement for reason attribution. Politicians tout their record on voting the same way for the same issues as evidence that they truly believe in that position. They are implicitly relying on the idea that someone motivated to act in a particular way for a particular reason is more likely to do so in the future. This reasoning is common in scholarship. For example, Saunders takes a strong position on the interpretation of statements made in times of crisis:

> Leaders may say and do things under the pressure of crisis decisionmaking that may not reflect their actual beliefs. Furthermore, stated beliefs may be merely post hoc justifications for action. Thus one cannot infer beliefs merely by observing leaders in crises. I therefore shift my primary measurement of causal beliefs to the prepresidential period, to show that presidents arrived in office with causal beliefs already in place (Saunders 2009, 135).

Here, Saunders is implicitly appealing to a form of the reason consistency idea. She is holding that beliefs (a central part of reasons for action) remain constant over time and can drive and explain behavior even in the face of contradictory contemporary statements. If the reason consistency idea is

rejected, then we would have no reason to think that presidents' beliefs before office are relevant to decisions made in office.

The consilience principle again proves a useful guide. If different actions are all consistent with a particular reason, then that reason attribution is to be preferred to one that is not consistent with those actions. However, as in the spaghetti example, changes in context should also be taken into account.

Context and reason attribution

Another type of evidence that might be used in a reason attribution is some property of the strategic context of the action performed. A contextual condition is some feature of an actor's situation, or environment, that plausibly provides him with an incentive to perform an action. This kind of evidence has the benefit of often being more easily accessible than statements made in a situation where there is no incentive for misrepresentation. Elster (2007, 61) raises the idea of using the 'objective interests' of an actor as a proxy for the actor's subjective motivation. However, as he also notes, this can suggest useful hypotheses, but does not dissolve the FPORA.

A recent example demonstrates both the importance of the strategic context in informing and motivating policy choices and the limitations of context as evidence of motive. Harvey (2011) uses extensive evidence, including private and public speeches and writings, to argue that the contextual conditions facing US president George Bush after 2001 would counterfactually have led 'president' Al Gore to make the same choices for the same reasons, culminating in going to war on Iraq in 2003. Harvey does not simply state the context and infer the reason from that. Instead, he carefully eliminates alternative motives for Bush's decisions, using a variety of types of evidence. What confidence we have in Harvey's conclusion derives from his use of different classes of evidence, leading to an inference to the most consilient reason.

Merely establishing that the context could provide an actor with a reason is only the first step in a reason attribution. Yet, sometimes work in IR relies too heavily on strategic context as evidence in reason attribution. The mere presence of a contextual condition cannot distinguish between reasons that are similarly supported by the presence of other contextual conditions. In addition, the actors in the strategic context, as described by scholars, may not be aware of that context. Some work does not adequately appreciate these points. One example is Simmons and Danner's analysis of the ICC (see above). They use the fact that a state is involved in a civil war as their only evidence that state's reason for joining the ICC was to credibly commit to abiding by a peace agreement in that civil war. They argue that states who have no other way to credibly commit to peace in a civil war join the

ICC in order to increase the costs of returning to war, and hence more credibly committing to a peace agreement. However, states in a civil war might also be involved in negotiations over trade deals (perhaps aimed at the post-war period) and want to use accession to the ICC as a bargaining chip in those negotiations.

Moreover, there can be more than one reason to perform the same action. That is, two actions that are of the same type can be motivated by two different reasons. One state in a civil war might want to credibly commit to peace using the ICC, but another might want to join the ICC because they think that this will enhance their chances of receiving external support in the civil war. The mere fact of being in a civil war cannot distinguish between these two possibilities.

Strategic context is thus useful for suggesting hypotheses of reasons, but by itself is only weak evidence. However, if a reason attribution is consistent with statements of reason *and* the strategic context, then it is to be preferred to a reason that is not. Here again, the principle of consilience subsumes our intuitions.

General comments on using evidence to determine motivation

Some scholars take a strong stand against the position that evidence of reasons can be used. For example, Jackson points out that 'there is no way to tell whether these reconstructed motives are the *real* motives possessed by the historical actors in question' (2002, 748) and Krebs and Jackson go further to deny that there is 'evidence [that] could even in principle clinch the case' (2007, 40). Similarly, Frieden argues that 'preferences are unobservable independent of outcomes' (1999, 48). Such a radical skepticism about evidence of reasons means resorting to an alternative response to the FPORA. Jackson also claims that 'given time, one can usually find evidence to support virtually any position in the documentary record' (2002, 748). This latter position is too strong. If you are accepting the possibility that you can use evidence to attribute reasons and you have a delimited time period and a clear group of actors you think were instrumental in the key decisions, the positions that receive support must be relatively limited and the weight of evidence may even be in only one direction, making inference to a historical reason more clear-cut.

Conclusion

Intentional explanation and hence reason attribution is pervasive in IR. If this is to continue as a major part of the way we understand the social world then the FPORA is an obstacle to the acceptability of some of our current research practices. This paper has highlighted the importance of

https://doi.org/10.1017/S175297191500007X Published online by Cambridge University Press

taking a stance on the basic issue facing anyone who desires to attribute reason; that we cannot introspect the reasons of others. At the very least, the knowledge claims you can consistently make vary with the type of response you take to the FPORA. If we take this problem seriously, our response to the problem should drive numerous other choices in terms of the status and extent of our work, the kinds of evidence we use, and how we use it.

Several broad lessons come out of the exploration of possible responses to the FPORA in this paper. First, explaining, or accounting for, known behavior using assumed reasons is a separate intellectual operation from using empirical evidence to establish reasons. Both should be evaluated on their own terms and, given practical limitations on space, time, and effort, should be the subject of separate research projects. This echoes Clarke and Primo's (2012) criticism of requiring formal models to be accompanied by some form of data analysis. However, it is also clear that assumption responses cannot stand alone; in order to know the actual reasons for action, we need to engage seriously with the empirical challenges of reason attribution.

Second, if your response to the FPORA is avoidance, then you are locating your explanatory power away from the level of the individual reason and so reason attribution is, strictly speaking, irrelevant. Therefore, while it might be interesting to speculate about reasons, explanations based on the avoidance response cannot be successfully challenged on the basis that they do not establish reasons, or eliminate alternative reasons. That said, there are potential costs to avoiding reason attribution, like missing out on sources of variation or novel theory generation.

Finally, there are multiple possible strategies when using empirical evidence in a reason attribution. Dividing up possible evidence into three types: statements, actions, and context, clarifies the particular issues involved in using each one. In addition, we can use the quality of consilience as a guide to choosing between reason attributions. The more types of evidence explained by or consistent with a reason attribution, the greater the credibility of that attribution. This is a superior guiding principle than existing exhortations to seek private evidence or to consider the context of documents.

Acknowledgments

The author thanks Robert Adcock, Davy Banks, Andrew Bennett, Alan Jacobs, Dianne Pfundstein, David Sylvan, Tristan Volpe, David Waldner, and participants at APSA 2013 for helpful comments. The paper benefited from the comments from three anonymous reviewers and the editors of *International Theory*, particularly Alex Wendt. The author is especially grateful to Robert Adcock for discussion and encouragement while the paper was in its early stages.

References

Adcock, Robert. 2003. "What Might it Mean to be an 'Interpretivist'?" *Qualitative Methods* 1(2): 16–18.

Andersen, Morten Skumsrud, and Iver B. Neumann. 2012. "Practices as Models: A Methodology with an Illustration Concerning Wampum Diplomacy." *Millennium: Journal of International Studies* 40(3):457–81.

Baldwin, Dare A., and Jodie A. Baird. 2001. "Discerning Intentions in Dynamic Human Action." *Trends in Cognitive Sciences* 5(4):171–78.

Baldwin, David. 1985. *Economic Statecraft*. Princeton, NJ: Princeton University Press.

Berard, Timothy. 1998. "Attributions and Avowals of Motive in the Study of Deviance." *Journal for the Theory of Social Behaviour* 28(2):193–213.

Bevir, Mark. 1999. *The Logic of the History of Ideas*. Cambridge: Cambridge University Press.

Bohl, Vivian, and Nivedita Gangopadhyay. 2014. "Theory of Mind and the Unobservability of Other Minds." *Philosophical Explorations* 17(2):203–22.

Bratman, Michael. 1981. "Intention and Means-End Reasoning." *The Philosophical Review* 90(2):252–65.

Broockman, David E. 2012. "The 'Problem of Preferences': Medicare and Business Support for the Welfare State." *Studies in American Political Development* 26(2):83–106.

Bruce, Steve, and Roy Wallis. 1983. "Rescuing Motives." *British Journal of Sociology* 34(1): 61–71.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph M. Siverson, and James D. Morrow. 2003. *The Logic of Political Survival*. Cambridge, MA: MIT Press.

Butler, Judith. 1990. "Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory." In *Performing Feminisms: Feminist Critical Theory and Theatre*, edited by Sue-Ellen Case, 270–82. Baltimore, MD: Johns Hopkins University Press.

Clarke, Kevin A., and David M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford: Oxford University Press.

D'Andrade, Roy G., and Claudia Strauss. eds. 1992. *Human Motives and Cultural Models*, vol. 1. Cambridge: Cambridge University Press.

Davidson, Donald. 1963. "Actions, Reasons and Causes." *Journal of Philosophy* 60:685–700.

Doenecke, Justus D. 1981. *The Diplomacy of Frustration: The Manchurian Crisis of 1931–1933 as Revealed in the Papers of Stanley K. Hornbeck*. Stanford, CA: Hoover Institution Press.

D'Oro, Giuseppina, and Constantine Sandis. 2013. *Reasons and Causes: Causalism and Anti-Causalism in the Philosophy of Action*. Basingstoke: Palgrave Macmillan.

Elster, Jon. 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Fearon, James D. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88(3):577–92.

—— 1995. "Rationalist Explanations for War." *International Organization* 49(3):379–414.

—— 1997. "Signaling Foreign Policy Interests: Tying Hands Versus Sinking Costs." *Journal of Conflict Resolution* 41(1):68–90.

Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97(1):75–90.

Frieden, Jeffry A. 1999. "Actors and Preferences in International Relations." In *Strategic Choice and International Relations*, edited by David A. Lake and Robert Powell, 39–76. Princeton, NJ: Princeton University Press.

Friedman, Milton. 1953. "The Methodology of Positive Economics." In *Essays in Positive Economics*, 3–16, 30–43. Chicago, IL: University of Chicago Press.

George, Alexander, and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Girard, Phillipe R. 2004. *Clinton in Haiti: The 1994 US Invasion of Haiti*. Basingstoke: Palgrave Macmillan.

Gunnell, John G. 2011. "Social Scientific Inquiry and Meta-Theoretical Fantasy: The Case of International Relations." *Review of International Studies* 37(4):1447–69.

Harvey, Frank P. 2011. *Explaining the Iraq War: Counterfactual Theory, Logic and Evidence*. Cambridge: Cambridge University Press.

Hayek, Friedrich A. 1980 [1952]. *The Counter-Revolution of Science*. Indianapolis, IN: Liberty Fund Inc.

Hirstein, William. 2005. *Brain Fiction: Self-Deception and the Riddle of Confabulation*. Cambridge, MA: MIT Press.

Hollen, Christopher Van. 1980. "The Tilt Policy Revisited: Nixon-Kissinger Geopolitics and South Asia." *Asian Survey* 20(2):339–61.

Hopf, Ted. 2007. "The Limits of Interpreting Evidence." In *Theory and Evidence in Comparative Politics and International Relations*, edited by Richard Ned Lebow and Mark Irving Lichbach, 55–84. Basingstoke: Palgrave-Macmillan.

Ikenberry, G. John. 2001. *After Victory: Institutions, Strategic Restraint, and the Rebuilding of Order After Major Wars*. Princeton, NJ: Princeton University Press.

Jackson, Patrick Thaddeus. 2002. "Jeremy Bentham, Foreign Secretary; or the Opportunity Costs of Neo-Utilitarian Analyses of Foreign Policy." *Review of International Political Economy* 9(4):735–53.

—— 2006. *Civilizing the Enemy: German Reconstruction and the Invention of the West*. Ann Arbor, MI: University of Michigan Press.

Jacobs, Alan. 2014. "Process-Tracing the Effects of Ideas." In *Process Tracing: From Metaphor to Analytic Tool*, edited by Andrew Bennett and Jeffrey T. Checkel, 41–73. Cambridge: Cambridge University Press.

Jansz, Jeroen. 1996. "Constructed Motives." *Theory and Psychology* 6(3):471–84.

Katznelson, Ira, and Barry Weingast. eds. 2005. *Preferences and Situations: Points of Intersection Between Historical and Rational Choice Institutionalism*. New York, NY: Russell Sage Foundation.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Krebs, Ronald, and Patrick Thaddeus Jackson. 2007. "Twisting Tongues and Twisting Arms: The Power of Political Rhetoric." *European Journal of International Relations* 13(1):35–66.

Krueger, Joel. 2012. "Seeing Mind in Action." *Phenomenology and Cognitive Sciences* 11: 149–73.

Laird, James D. 2007. *Feelings: The Perception of Self*. Oxford: Oxford University Press.

Lebovic, James, and Erik Voeten. 2006. "The Politics of Shame: The Condemnation of Country Human Rights Practices in the UNHRC." *International Studies Quarterly* 50(4):861–88.

Lebow, Richard Ned. 2010. *Why Nations Fight: Past and Future Motives for War*. Cambridge: Cambridge University Press.

Lichtenstein, Sarah, and Paul Slovic. eds. 2006. *The Construction of Preference*. Cambridge: Cambridge University Press.

Lipton, Peter. 1991. *Inference to the Best Explanation*. New York, NY: Routledge.

Lisieux, St Thérèse of. 2005. *The Story of a Soul (L'Histoire d'une Ame): The Autobiography of St. Thérèse of Lisieux*. Translated by Thomas Taylor. http://www.gutenberg.org/ebooks/16772.

Mackie, Gerry. 1998. "Are All Men Liars?." In *Deliberative Democracy*, edited by Jon Elster, 69–96. Cambridge: Cambridge University Press.

Mahoney, James. 2008. "Toward a Unified Theory of Causality." *Comparative Political Studies* 41(4/5):412–36.
March, James G., and Johan P. Olsen. 1998. "The Institutional Dynamics of International Political Orders." *International Organization* 52(4):943–69.
Martin, John Levi. 2011. *The Explanation of Social Action*. Oxford: Oxford University Press.
Mayhew, David R. 1974. *Congress: The Electoral Connection*. New Haven, CT: Yale University Press.
McGrew, Timothy. 2003. "Confirmation, Heuristics, and Explanatory Reasoning." *British Journal for Philosophy of Science* 54:553–67.
Mercer, Jonathan. 2012. "Audience Costs are Toys." *Security Studies* 21(3):398–404.
Mills, C. Wright. 1940. "Situated Actions and Vocabularies of Motive." *American Sociological Review* 5(6):904–13.
Moon, J. Donald. 1975. "The Logic of Political Inquiry: A Synthesis of Opposed Perspectives." In *Handbook of Political Science*, vol. 1, edited by Fred I. Greenstein and Nelson W. Polsby, 131–228. Reading, MA: Addison Wesley.
Morgenthau, Henry J. 1993 [1948]. *Politics Among Nations: The Struggles for Power and Peace*. Boston, MA: McGraw-Hill.
Nexon, Daniel H. 2009. *The Struggle for Power in Early Modern Europe: Religious Conflict, Dynastic Empires, and International Change*. Princeton, NJ: Princeton University Press.
Renshon, Jonathon. 2009. "When Public Statements Reveal Private Beliefs: Assessing Operational Codes at a Distance." *Political Psychology* 30(4):649–61.
Ringmar, Erik. 2014. "The Search for Dialogue as a Hindrance to Understanding: Practices as Inter-Paradigmatic Research Program." *International Theory* 6(1):1–27.
Rosato, Sebastian. 2014. "The Inscrutable Intentions of Great Powers." *International Security* 39(3):48–88.
Rubinstein, Ariel. 1991. "Comments on the Interpretation of Game Theory." *Econometrica* 59(4):909–24.
Saunders, Elizabeth. 2009. "Transformative Choices: Leaders and the Origins of Intervention Strategy." *International Security* 34(2):119–61.
Schelling, Thomas. 1978. *Micromotives and Macrobehavior*. New York, NY: Norton.
—— 1980. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
Schilbach, Leonhard, Bert Timmermans, Vasudevi Reddy, Alan Costall, Gary Bente, Tobias Schlicht, and Kai Vogeley. 2013. "Toward a Second-Person Neuroscience." *Behavioral and Brain Sciences* 36(4):393–414.
Schultz, Kenneth A. 2012. "Why We Needed Audience Costs and What We Need Now." *Security Studies* 21(3):369–75.
Simmons, Beth A., and Alison Danner. 2010. "Credible Commitments and the International Criminal Court." *International Organization* 64(2):225–56.
Skinner, Quentin. 2002. *Visions of Politics, Volume 1: Regarding Method*. Cambridge: Cambridge University Press.
Snyder, Jack, and Erica D. Borghard. 2011. "The Cost of Empty Threats: A Penny, Not a Pound." *American Political Science Review* 105(3):437–56.
Taylor, Charles. 1971. "Interpretation and the Sciences of Man." *The Review of Metaphysics* 25(1):3–51.
—— 1993. "To Follow a Rule." In *Bourdieu: Critical Perspectives*, edited by Craig Calhoun, Edward LiPuma, and Moishe Postone, 45–60. Cambridge: Polity Press.
Trachtenberg, Marc. 2012. "Audience Costs: An Historical Analysis." *Security Studies* 21(1):3–42.
Weber, Max. 1978. "Basic Sociological Terms." In *Economy and Society: An Outline of Interpretive Sociology*, 2 vols, edited by Guenther Roth and Claus Wittich, 3–63. Berkeley, CA: University of California Press.

Wendt, Alexander. 2001. "Driving with the Rearview Mirror: On the Rational Science of Institutional Design." *International Organization* 55(4):1019–49.
Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Chichester: Wiley-Blackwell.
Zahavi, Dan, and Shaun Gallagher. 2008. "The (In) Visibility of Others: A Reply to Herschbach." *Philosophical Explorations* 11(3):237–44.