Episteme (2024), 1–11
doi:10.1017/epi.2023.58

episteme

ARTICLE

Assertions: Deterrent or Handicap? A Reply to Graham (2020)

Justin P. Bruner

State University of New York at Buffalo, Buffalo, NY, USA
Email: jbruner@buffalo.edu

(Received 13 August 2022; revised 17 July 2023; accepted 17 October 2023)

Abstract

According to one influential tradition, to assert that p is to express a belief that p. Yet how do assertions provide strong evidence for belief? Philosophers have recently drawn on evolutionary biology to help explain the stability of assertive communication. Mitchell Green suggests that assertions are akin to biological handicaps. Peter Graham argues against the handicap view and instead claims that the norms of assertion are deterrents. Contra Graham, I argue that both mechanisms may play a role in assertive communication, although assertions as deterrents will often fail to provide strong evidence for belief.

Key words: Assertion; handicap principle; animal communication; social epistemology; gametheory; social norms

1. Speech acts and stability

According to one influential tradition, to assert that p is to express a belief that p. Yet this appears to require a certain level of sincerity on the part of the speaker. If the interests of our interlocutors are not perfectly aligned, a speaker may have incentive to assert p even though she does not in fact believe that p. In such a case, can we still claim that the speaker is asserting that p? To address this query, many have emphasized that assertions provide *evidence* for our beliefs, and in the case of insincerity (where one asserts p even though one does not believe that p), assertions provide misleading evidence of the speaker's underlying beliefs. Yet how, precisely, do assertions play this evidentiary role?

Mitchell Green (2009) provides an innovative response. Drawing on evolutionary theory and game theory,[1] Green contends that assertions provide ample evidence for our beliefs because assertions are handicaps. What does Green mean by this? A major question in the study of animal signaling is how honest communication is possible when the interests of sender and receiver are not perfectly aligned. If the sender sometimes has incentive to mislead his counterpart, then shouldn't receivers for the most part ignore senders? And if receivers discount incoming signals, why would

[1]For more on the use of evolution and game theory in the philosophy of language, see Lewis (1969), Skyrms (1996), Huttegger (2007), Zollman (2011).

their counterpart ever bother to send a message? One solution comes in the form of costly signaling theory, which stipulates that communication is possible when some signals are costly to produce. These "handicaps" are thought to stabilize communication in nature. Moreover, signal production cost is not the same for all senders: "dishonest" signalers pay a significantly larger cost than "honest" signalers to produce the signal.

In light of this cost differential, only "honest" senders are incentivized to signal.[2] This reasoning has assisted biologists in making sense of seemingly perplexing signaling behavior, such as the peacock's tail: while costly to all male peacocks, the tail is presumably more burdensome to males of low genetic quality. Thus, the presence or absence of an ornate and colorful tail is indicative of the male's underlying genetics.

Green argues that assertions are governed by a similar kind of logic. When one asserts that p, one takes on a liability, and thus, assertions "carry a cost" (157). Specifically, the norms governing speech acts result in a loss of credibility in certain circumstances. A norm of assertion might demand we only assert if we believe p on very good evidence. When the speaker flouts this norm, they risk a loss of credibility. The threat of such a loss is what enables us to "discern a connection between speech acts and handicaps" (153). Loss of credibility ensures speakers only assert that p when they believe that p on very good evidence.

In his 2020 article "Assertions, handicaps and social norms," Peter Graham (2020) rightly questions the connection between speech acts and handicaps. Graham notes that there are important dissimilarities between canonical handicaps (the peacock's tail) and assertions. In the case of the peacock, signaling costs are paid at equilibrium: "honest" senders possess an ornate (and costly) tail. This is not true of assertions. A cost (loss of credibility) is only paid outside of equilibrium by those "dishonest" interlocutors. Furthermore, Graham notes that handicaps "only reliably signal the quality they waste." To use a different example, a suburbanite may attempt to signal their wealth by wasting their wealth on an utterly impractical, but nonetheless expensive, lifted pickup truck. Because an 80,000-dollar vehicle is prohibitively costly to those of modest means, our suburbanite reliably signals their wealth when they waste their wealth on what is essentially a large toy used to haul groceries from Costco twice a month. Assertions – on the other hand – do not appear to work in this fashion: by wasting credibility one does not thereby signal credibility.

For these reasons, Graham rejects Green's claim that assertions are handicaps, opting for an alternative mechanism: deterrents. Simply put, a deterrent is a "cost that a dishonest signaler is likely to pay if caught making a dishonest signal" (365). As Graham notes, this seemingly straightforward mechanism requires three elements: first, it must be possible to determine whether the signal is "honest" or not. For deterrents to be effective, verification is needed. Furthermore, the recipient of the signal must have a means of punishing "dishonest" signalers. Finally, the punishment must swamp any benefits associated with deception. When these three requirements are met, deterrents may play a crucial role in stabilizing[3] animal signaling systems. Yet how do deterrents stabilize assertive communication? Recall that Green believes assertions carry a "liability" and thus credibility is lost if one flouts the norms of assertion. Those who defy the norms of assertion by – for instance – asserting that p when they do not believe that p,

[2]See Zahavi (1975) for an informal statement of this mechanism and Grafen (1990) for a careful game-theoretic treatment. The economist Michael Spence came to a similar conclusion nearly two decades prior to Grafen's paper and received the Nobel Prize for his work on costly signaling.

[3]See, for example, the case of status badges among sparrows as discussed in Graham.

https://doi.org/10.1017/epi.2023.58 Published online by Cambridge University Press

suffer a loss in credibility. Hence, only those "dishonest" speakers pay a cost, and only if their norm violation is uncovered by their interlocutor. Contra Green, it appears that "norms of assertion are deterrents" (359).

2. Can deterrents provide strong evidence for our beliefs?

The honesty-stabilizing mechanism discussed in Green – whereby speakers incur a loss of credibility should they flout the norms of assertion – is incorrectly identified by Green as involving a handicap. Instead, as Graham persuasively argues, the mechanism described and discussed in Green is best seen as a kind of deterrent. This mislabeling correction aside, Green's motivating question – can assertions provide evidence (and moreover, strong evidence) for our beliefs? – has yet to be satisfactorily addressed. In this section, I suggest the answer is a not so satisfying "it's complicated." With the help of a simple game-theoretic model, I contend that deterrents undergird a signaling system where speakers often flout the norms of assertion. I then draw on a measure of evidential support from Bayesian confirmation theory and show that while assertions always provide evidence for our beliefs at the deterrent backed equilibrium, in many cases the evidential support is weak or nearly non-existent. In what follows, we motivate and introduce our strategic scenario of interest, which is then modelled and analyzed game-theoretically. For simplicity, we consider just violations of the norms of assertion that involve insincerity. A similar analysis can be conducted for violations where individuals assert that p even though they believe that p but not on the basis of very good evidence.[4]

2.1. To bake, or not to bake?

Alf and Betty are good friends and are considering a joint business venture. Alf recently graduated from culinary school and is interested in opening a bakery. He has all the know-how but cannot open the store without a loan. Betty knows nothing of baking but was recently bequeathed a large sum of money from her rich aunt and is considering whether to use the funds to finance Alf's bakery. Because she trusts Alf's judgment, Betty is willing to funnel her inheritance into the bakery if and only if Alf believes the market is favorable. Regardless of the state of the market, Alf prefers the injection of cash from Betty.

Because he is an industry insider, it is not too difficult for Alf to determine the condition of the market, and we assume for simplicity that he never errs in his assessment.[5] With probability q the business environment is favorable (and thus Alf believes that the market is favorable); with probability $1 - q$ the business environment is unfavorable. In the latter case, although Alf does not believe that p ("The market is favorable"), he nonetheless benefits if Betty comes to believe that p. Thus, Alf might assert that p, in

[4]I strongly suspect that such an analysis will come to the same conclusion we arrive at in this paper. Namely, speakers will routinely flout the norms of assertion at the deterrent backed equilibrium and assertions as deterrents often provide little in the way of evidential support. My reason for thinking this is as follows: if individuals gain socially from contributing to the epistemic commons, there is a temptation to assert that p even though one's belief in p is not on the basis of very good evidence. Thus, there is a conflict of interest between speaker and hearer. Speaker would prefer to flout the norms (assuming they don't get caught), while hearer prefers the speaker always adheres to the norms of assertion. This conflict of interest is all we need to get to the partially informative deterrent backed equilibrium.

[5]That is, Alf believes that the market is (un)favorable if and only if the market actually is (un)favorable.

https://doi.org/10.1017/epi.2023.58 Published online by Cambridge University Press

(a)

	C (Credulous)	I (Investigate)
A (Always conform)	1, 1	1, 1 – m
F (Flout only when interests do not coincide)	1, 1	1, 1 – m

(b)

	C (Credulous)	I (Investigate)
A (Always conform)	0, 1	0, 1 – m
F (Flout only when interests do not coincide)	1, 0	– s, 1 – m – k

Figure 1. Game between Alf and Betty where (a) the interests of the two parties do not conflict and (b) the interest of the two parties do conflict.

the hopes that Betty comes to believe that p. Considering this scenario from a game-theoretic perspective, Alf can be seen as having two strategies: always conform to the norms of assertion (and thus do not assert that p in the above case), or flout the norms when the interests of interlocutors come into conflict (and thus assert that p in the above case). Betty, on the other hand, can respond to Alf's assertion that p in one of two ways. She can credulously take the assertion that p to indicate that Alf believes that p, or she can investigate whether Alf violated the norms of assertion. If Alf is shown to have violated the norms of assertion, Betty does not adopt the belief that p. Moreover, she punishes Alf and Alf consequently suffers a credibility loss.

For ease of understanding, we break this scenario down into two cases (Figure 1) and eventually combine them to form one "master" game (Figure 2). Figure 1a shows the payoffs for both Alf and Betty when their interests coincide (favorable market). Alf believes that p and asserts that p. Regardless of whether she is credulous or not, Betty adopts the belief that p and the bakery is a success (payoff of one).[6] In Figure 1b we have the payoffs for Alf and Betty when their interests conflict (unfavorable market). In this case, Alf does not believe that the market is favorable. Nonetheless, he benefits if Betty comes to believe that p. If he conforms to the norms of assertion, Alf does not assert that the market is favorable and thus Betty does not come to believe that the market is favorable. This is bad for Alf (payoff of zero), but good for Betty (payoff of one). If he flouts the norms and Betty is credulous, then Betty adopts the belief that the markets are favorable, and she invests. This outcome is bad for Betty (payoff of zero) but good for Alf (payoff of 1). Finally, if Betty decides to interrogate Alf's assertion, she never comes to believe that the market is favorable.[7] However, interrogation is costly (m) and moreover, if Alf flouts the norms, Betty pays an additional cost to punish Alf for his transgression (c) and Alf suffers a loss of credibility (s).

We combine the two tables from Figure 1 to create the game displayed in Figure 2. What kind of behavior should we expect in this strategic scenario, and moreover, will

[6]That is, we assume Betty's investigation is perfectly informative (see footnote 7 as well).

[7]This, of course, is not particularly realistic to assume, but we do so to ensure that deterrence has the best possible chance of grounding assertive communication.

https://doi.org/10.1017/epi.2023.58 Published online by Cambridge University Press

	Credulous	Investigate
A (Always conform)	q, 1	q, 1 – m
F (Flout only when interests do not coincide)	1, q	q + (1-q)(– s), 1 – m – k(1-q)

Figure 2. Game between Alf and Betty where with probability q the interests of the parties do not conflict.

Alf adhere to the norms of assertion at equilibrium? Let's begin by considering the case where Alf adheres to the norms and Betty interrogates. This arrangement is not stable because at least one person (Betty) has incentive to change their behavior. Given that Alf is adhering to the norms, Betty does best not to interrogate. Yet once Betty switches strategies, then we're left in a similar situation: at least one individual (Alf) has incentive to change their behavior. Once Betty no longer decides to interrogate her interlocutor, Alf does best to flout the norms of assertion. It appears there is no stable equilibrium in this strategic scenario. Alf and Betty chase each other around in a circle from now until the end of time. Deterrents don't appear to have helped much.

Fortunately, there does exist a stable Nash equilibrium where Alf *sometimes* abides by the norms of assertion. Specifically, there exists a mixed-strategy Nash equilibrium wherein Alf plays A with probability $1 - m/[(1 - q)\ (1 - c)]$, meaning he is less likely to conform to the norms when the cost of inspection (m) increases or either the probability of a favorable market (q) or cost of punishing (k) goes up. Betty plays I with probability $1/(1 + s)$, meaning the probability of inspection is purely a function of how severe Alf's punishment is. She's less likely to inspect the more severe the punishment.

Deterrents are thus an imperfect means of stabilizing assertive communication: with some non-trivial probability, Alf will assert that p when he does not in fact believe that p. This realization on its own is not too surprising. As Graham notes, punishment is often fickle and comes with its own stability problems.[8] Furthermore, as we mentioned above, there are no equilibria involving just pure strategies. If Alf always conforms to A, then Betty does best to not inspect, which incentivizes Alf to flout the norms. It makes sense that we arrive at a mixed equilibrium where Alf sometimes adheres and sometimes flouts the norms. Yet, as we shall see, assertions at this deterrent backed equilibrium often provide very weak evidence – and in some cases almost no evidence at all – for our beliefs.

2.2. Assertions and evidential support

Recall that for Green, assertions express belief. Yet Betty cannot directly observe Alf's beliefs, which motivates Green's central question: "how it is we provide strong evidence for our beliefs – how do we 'express' our beliefs by asserting?" (Graham, 349). Green's answer is that assertions are strong evidence for belief because they are handicaps. Graham's correction is that assertions are in fact deterrents. Yet in many quotidian cases, assertions as deterrents do not provide especially strong evidence for our beliefs, or so I shall argue. So, even if Graham is correct to emphasize the importance of deterrents, we are left with a rather tenuous connection between assertions and beliefs (in the

[8]See footnote 10 of Graham.

https://doi.org/10.1017/epi.2023.58 Published online by Cambridge University Press

next section I suggest that Graham was perhaps too quick to discard handicaps and that handicaps – or at least a closely related "handicap-like" mechanism – can in fact provide strong evidence for beliefs).

To appreciate this, consider a widely known, albeit flat-footed, explication of the concept of evidence. Simply, e is evidence for some hypothesis H (in our case, the hypothesis that Alf believes that the market is favorable) when pr(H | e) > pr(H). That is, e confirms H when e makes H more likely to be true. Returning to our friends Alf and Betty, the market is favorable with probability q, meaning with probability q Alf believes the market is favorable (i.e., pr(H) = q). Now, consider the probability that Alf believes that the market is favorable conditional on Alf asserting that the market is favorable at the deterrent backed mixed-strategy Nash equilibrium. In other words, if Alf asserts that the market is favorable, what is the likelihood that Alf *actually believes this* (at equilibrium)? Using Bayes' rule, we see that this probability is $1 - m/[m + (1 - c)q]$. Importantly, this value will always be greater than p(H) at the deterrent backed Nash equilibrium.[9] So, even though Alf might frequently flout the norms of assertion at equilibrium, assertions as deterrents are evidence for our beliefs.

Yet, do assertions as deterrents provide *strong* evidence for our beliefs? Do deterrents really succeed where handicaps have supposedly failed? To address this question, let's begin with a somewhat extreme, but nonetheless informative, numerical example. If q = 0.66, m = 0.2, and c = 0.4, then at the deterrent backed equilibrium, the probability that Alf believes that conditions are favorable (Pr(H)) is 0.66, and the probability that Alf believes that the market is favorable on the condition that Alf asserts that the market is favorable is 0.6644. So while Pr(H | e) is greater than Pr(H), Alf's assertion does not seem to provide much in the way of evidential support for his beliefs: his assertion only shifts the probability that he believes that the market is favorable from 0.66 to 0.6644. How common is it for assertions to provide minimal or almost no evidential support for our beliefs at the deterrent backed equilibrium? To more rigorously explore this question, we draw on a specific formal measure of evidential support known as the ratio measure.[10] According to this measure, the degree to which e supports hypothesis H depends on a comparison between the probability of the hypothesis given the evidence, and the unconditional probability of the hypothesis. This gets us the following ratio: Pr(H | e)/Pr(H). If the evidence does not lend any support to the hypothesis, then (Pr(H | e) = Pr(H)), and this ratio is equal to one. We take the logarithm of this ratio to ensure that the value of the measure is zero when the evidence provides no support for the hypothesis. So, the degree of evidential support is log(Pr(H | e)/Pr(H)).

With this measure in hand, we determine the degree of evidential support provided by assertions at the deterrent backed equilibrium for various parameter values (Figure 3). Note that the level of support decreases as q – the probability interests coincide (i.e., the probability that market is favorable) – goes up. For low values of q (say, q = 0.01) the conditional probability is not much greater than Pr(H) in absolute terms (0.0291 versus 0.01[11]), but the ratio of these terms is quite substantial (2.91). On the other hand, when q is 0.5, the difference between pr(H) and pr(H | e) is more substantial (but still quite modest: 0.6 versus 0.5) while the ratio is only 1.2.

[9]Specifically, p(H | e) is greater than p(H) when q is less than $1 - m/(1 - c)$. The deterrent backed mixed NE exists when this very condition is met, so assertions always are evidence of our beliefs at the deterrent-backed equilibrium.

[10]For more on this measure, as well as alternatives, see Fitelson (2001).

[11]We are assuming (as we do in Figure 3) that m = 0.2 and c = 0.4.

https://doi.org/10.1017/epi.2023.58 Published online by Cambridge University Press

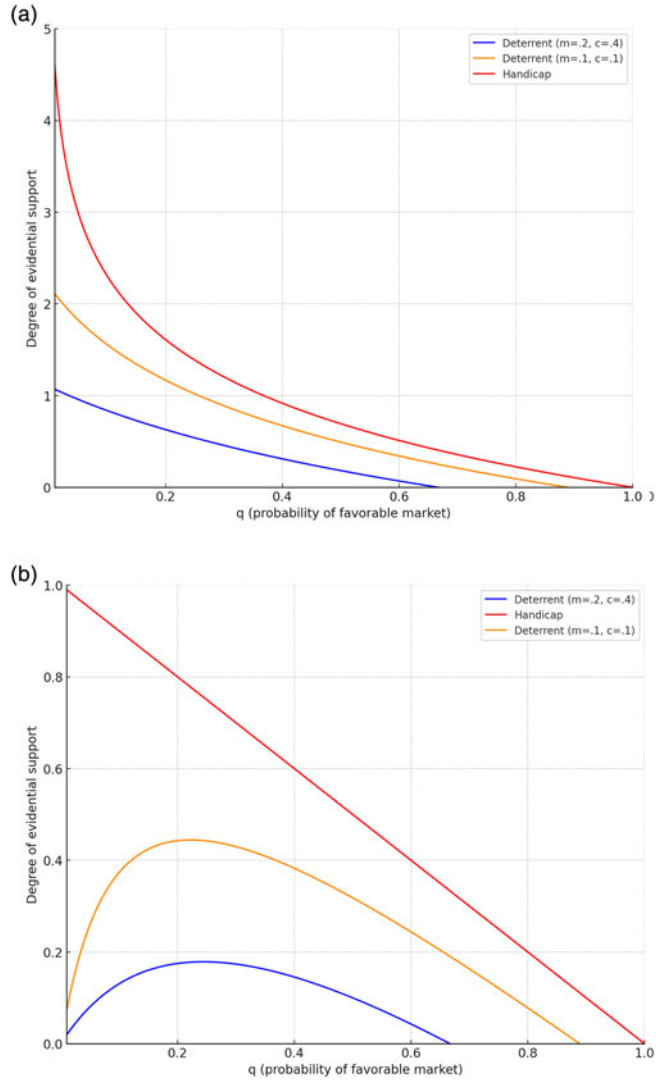

Figure 3. (a) Degree of evidential support according to the ratio measure [log(p(H | e)/p(H))] at the handicap and deterrent backed equilibria. (b) Degree of evidential support according to the difference measure [p(H | e) – p(H)] at the handicap and deterrent backed equilibria.

For the sake of comparison – and to give some additional context to these numbers – recall the handicap backed signaling system discussed in Green. We determine the degree of evidential support provided by assertions at this equilibrium.[12] Ignoring Graham's criticisms detailed in the previous section, at a handicap backed signaling

[12]In other words, we are considering what is often referred to as the "separating equilibrium," where costly signals ensure perfect information transfer between sender and receiver. In Section 4 we note how a partially informative equilibrium – similar to the deterrent backed equilibrium – is also possible when

https://doi.org/10.1017/epi.2023.58 Published online by Cambridge University Press

system Alf asserts that the market is favorable if and only if he in fact believes that the market is favorable, meaning pr(H | e) is equal to one. If the market is favorable with probability 0.66, then the degree of evidential support provided by Alf asserting that the market is favorable at the handicap backed signaling system is log(1/0.66) or 0.4155. Recall that the measure of support at the deterrent backed equilibrium is 0.0291. So, Green's signaling system is quite effective, and assertions at this handicap backed equilibrium provide much stronger evidence for our beliefs than assertions at the deterrent backed equilibrium.[13] The level of evidential support at the deterrent backed equilibrium is not a function of the ratio measure. As Figure 3b makes clear, similar results hold if we instead use a measure of evidential support sometimes referred to as the difference measure, where the degree of support is simply p(H | e) – p(H).

3. Assertions can be handicap-like

Deterrents may have a difficult time stabilizing assertive communication, or so I have argued. However, this difficulty may be the result of one crucial assumption. Namely, that it is the listener who doles out punishment. Our discussion has assumed that it is the responsibility of the listener to both determine whether a signal is "dishonest" as well as sanction offending parties. Yet, what if sanctions are not at the behest of some other party, but instead originate from the speaker themselves. Internal sanctions of this kind are not too difficult to imagine. For instance, Alf might feel a pang of guilt when he asserts that the markets are favorable when he in fact knows they are in dire straits. As was the case with external punishment, internal sanctions can be effective, but they must be large enough to outweigh the benefits of deception. If this latter point holds, then internal sanctions can stabilize a signaling system wherein Alf asserts that p only when Alf believes that p. Flouting the norms of assertion would be too onerous once internal sanctions are accounted for. Relatedly, if Betty knows that Alf is guilt-prone or has otherwise internalized the norms of assertion, then she can be sure that Alf believes that p when he asserts that p.

Internal sanctions, then, are an excellent means of stabilizing assertive communication. This mechanism, however, bears a striking resemblance to handicaps. First, recall that handicaps can stabilize a signaling system because "dishonest" signalers pay a large cost to produce the signal. If the cost is high enough, deception is discouraged. Internal sanctions can be seen in a similar light: when Alf signals dishonestly, he will feel guilt or shame. These psychological costs are paid whenever Alf produces a dishonest signal and – when sufficiently high – discourage deception.[14] This brings us to our second crucial point, that both handicaps and internal sanctions can support an equilibrium involving no deception. Hence, the signaling system stabilized by internal sanctions more closely resembles the handicap backed equilibrium than the deterrent backed mixed-strategy Nash equilibrium from the previous section.

handicaps are at play. Deterrents – as introduced and discussed by Graham – do not allow for a separating equilibrium.

[13]This claim of course does not hold for all values of m and c. As m and c go to zero the probability Alf plays A at the mixed-strategy Nash equilibrium approaches 1, meaning the deterrent backed equilibrium strongly resembles the handicap backed equilibrium.

[14]These emotions are not experienced by Alf should he signal honestly, meaning there is differential signal cost (a crucial feature of handicaps).

https://doi.org/10.1017/epi.2023.58 Published online by Cambridge University Press

So, maybe we were too quick to discard handicaps. Not all assertions are handicaps – this much is obvious – but perhaps *some* assertions provide strong evidence for the speaker's beliefs because these assertions are handicap-like. I say handicap-like because Graham would most likely be reluctant to refer to internal sanctions as handicaps because they fail what he views as a crucial test. For Graham, a handicap (or more specifically, what he refers to as a quality-handicap) must waste what it attempts to show.[15] Our suburbanite wastes wealth and purchases an expensive pickup truck in order to signal wealth. It is clear that internal sanctions are not handicaps in this strict sense: what Alf attempts to show (a belief) is not what is wasted.

We now consider one final reason why internal sanctions are more akin to handicaps than deterrents. Following ornithologists William Searcy and Stephen Nowicki, it is common practice in costly signaling theory to distinguish between two broad classes of signal costs. On the one hand, costs can be receiver-independent, meaning they are "imposed regardless of whether or how receivers respond" (Searcy and Nowicki 2005, as cited in Fraser 2012). Handicaps typically involve receiver-independent costs: an ornate tail is a burden no matter the response elicited from the female peacock. Deterrents, as described and discussed by Graham, do not involve receiver-independent costs. Instead, whether Alf suffers a loss of credibility is a function of how Betty (the receiver of the signal) responds. Deterrents involve receiver-dependent costs.

Another example of a signal with receiver-dependent costs are the so-called vulnerability signals – signals that expose the sender to potential harm and place them at risk. The breast-to-breast display of the male fulmar is such a signal (Enquist *et al.* 1985). In adversarial settings, closing the distance between yourself and a rival (as the breast-to-breast display does) is not on its own costly. However, it places one in an especially vulnerable position should your counterpart attack, meaning the cost associated with the display is a function of (among other things) how the receiver responds. Perhaps not surprisingly, game-theoretic analysis of vulnerability signals indicates these signals underwrite a mixed-strategy Nash equilibrium similar to the deterrent backed equilibrium uncovered in Section 2 (Adams and Mesterton-Gibson 1995; Bruner *et al.* 2017). In other words, deterrents and vulnerability signals – both signal-types that involve receiver-dependent costs – are viable, albeit imperfect, means of thwarting deception. Finally, it should be clear that internal sanctions involve receiver-independent costs. The psychological costs incurred by Alf are merely a function of his behavior and are not influenced by Betty's response. So, once again, internal sanctions more closely resemble handicaps than deterrents.

4. Concluding remarks: assertive communication and the Robinson Crusoe fallacy

We have argued that assertions as deterrents need not provide strong evidence for beliefs. At the deterrent backed mixed-strategy Nash equilibrium, senders do not always adhere to the norms of assertion. This means that assertions often provide minimal evidential support at equilibrium. Our results hinge on the fact that punishment for norm

[15]As mentioned in Section 1, Graham also thinks assertions are not handicaps because he mistakenly believes that when handicaps are at play, signals must be costly at equilibrium. This is false, as Michael Lachmann, Szabolcs Szamado, and Carl Bergstrom show in a famous paper (Lachmann *et al.* 2001). Modifying a simple model meant to capture Zahavi's handicap principle, they show that at equilibrium it is possible for all signalers to "signal their true quality with zero cost" (13190).

https://doi.org/10.1017/epi.2023.58 Published online by Cambridge University Press

violations is administered by the hearer, and the hearer is an expected utility maximizer who must spend time, energy and resources verifying the claims of her interlocutor and administering punishment when appropriate.[16] If instead punishment and monitoring is outsourced to an all-knowing third-party who is concerned only with minimizing norm violations – their own utility be damned! – then speakers will never flout the norms and assertions will thus provide strong evidence for beliefs. Unfortunately, epistemic communities can rarely, if ever, rely upon such selfless agents. Instead, whoever is ultimately responsible for verification and punishment will have to determine how much of their scarce time, energy, and resources to devote to the task, and this decision will in turn be informed by how likely it is that speakers flout the norm (there is no point in monitoring when norm violations are a rare occurrence). Yet the likelihood of a norm violation is not "exogenous but part of the (equilibrium) strategy of a rational opponent," and to treat the probability of a norm violation as fixed is to commit what political scientist George Tsebelis calls the Robinson Crusoe fallacy (Tsebelis 1989). It is only when we are mindful of this fallacy, and do not mistake what is a game of strategy for a decision problem, do we uncover the kind of partial enforcement we see at the mixed-strategy Nash equilibrium of Section 2.

Luckily, deterrents are not the only game in town, or so I have argued. At their best, handicap-like mechanisms (such as internal sanctions) do not result in a mixed-strategy Nash equilibrium, but instead a signaling system where the sender always adheres to the norms of assertion. Handicap-like mechanisms have another benefit compared to deterrents: at equilibrium, communication does not involve costs for either speaker or listener. This is in stark contrast to deterrents, as the deterrent backed equilibrium can be quite costly for both speaker (they may incur a loss of credibility) as well as listener (monitoring and punishment costs). In other words, the expected costs for both speaker and listener are substantial at the deterrent backed equilibrium; the expected costs for speaker and listener are zero when handicap-like mechanisms work as they should. A final worry one may have about internal sanctions is that they may prove ineffective when the level of guilt or shame felt by the speaker is not sufficiently high to prevent the speaker from flouting the norms of assertion. This is a legitimate worry, but one which is mediated by the fact that game theorists have recently shown how "cheaper than costly" handicaps can result in a partially informative signaling system, similar to the deterrent backed mixed-strategy equilibrium.[17] So, even when internal sanctions are mild, we have reason to think that handicap-like mechanisms do no worse of a job than deterrents at stabilizing assertive communication.

Overall, while handicaps have many advantages over deterrents, deterrents surely play a role in our day-to-day interactions. My goal in this paper is not to discount deterrents entirely, but establish that this mechanism may not always be able to play the role envisioned by Green and Graham. Yet, as I hope has been made clear, the failure of deterrents is by no means catastrophic. Alternative mechanisms, such as handicaps, can lend a hand.[18]

[16]This is in line with Graham's description of deterrents provided on page 356 of Graham and briefly summarized in Section 1 of this paper.

[17]Elliott Wagner, Kevin Zollman, Simon Huttegger, and Carl Bergstrom have published a number of great articles on the evolutionary significance of "cheaper than costly" signals. See Wagner (2013), Zollman *et al.* (2013), and Huttegger and Zollman (2010).

[18]Thanks to Peter Graham and Hannah Rubin for helpful comments and conversation.

https://doi.org/10.1017/epi.2023.58 Published online by Cambridge University Press

References

Adams E. and Mesterton-Gibson M. (1995). 'The Cost of Threat Displays and the Stability of Deceptive Communication.' *Journal of Theoretical Biology* **175**(4), 405–21.

Bruner J., Brusse C. and Kalkman D. (2017). 'Cost, Expenditure and Vulnerability.' *Biology and Philosophy* **32**, 257–75.

Enquist M., Plane E. and Roed J. (1985). 'Aggressive Communication in Fulmars (Folmarus Glacialis) Competing for Good.' *Animal Behavior* **33**, 1007–20.

Fitelson B. (2001). *Studies in Bayesian Confirmation Theory*. Dissertation.

Fraser B. (2012). 'Costly Signaling Theories: Beyond the Handicap Principle.' *Biology and Philosophy* **27**, 263–78.

Grafen A. (1990). 'Biological Signals as Handicaps.' *Journal of Theoretical Biology* **144**(4), 517–46.

Graham P. (2020). 'Assertions, Handicaps and Social Norms.' *Episteme* **17**, 349–63.

Green M. (2009). 'Speech Acts, the Handicap Principle and the Expression of Psychological States.' *Mind and Language* **24**(2), 139–63.

Huttegger S. (2007). 'Evolutionary Explanations of Indicatives and Imperatives.' *Erkenntnis* **66**, 409–36.

Huttegger S. and Zollman K.J.S. (2010). 'Dynamic Stability and Basins of Attraction in the Sir Philip Sidney Game.' *Proceedings of the Royal Society B* **277**, 1915–22.

Lachmann M., Számádo S. and Bergstrom C. (2001). 'Cost and Conflict in Animal Signals and Human Language.' *Proceedings of the National Academy of Sciences USA* **98**(23), 13189–94.

Lewis D. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Searcy W. and Nowicki S. (2005). *The Evolution of Animal Communication*. Princeton, NJ: Princeton University Press.

Skyrms B. (1996). *The Evolution of the Social Contract*. Cambridge, MA: Cambridge University Press.

Tsebelis G. (1989). 'The Abuse of Probability in Political Analysis: The Robinson Crusoe Fallacy.' *American Political Science Review* **83**(1), 77–91.

Wagner E. (2013). 'The Dynamics of Costly Signaling.' *Games* **4**(2), 163–81.

Zahavi A. (1975). 'Mate Selection: A Selection for a Handicap.' *Journal of Theoretical Biology* **53**(1), 205–14.

Zollman K. (2011). 'Separating Directives and Assertions Using Simple Signaling Games.' *Journal of Philosophy* **108**(3), 158–69.

Zollman K.J.S., Bergstrom C.T. and Huttegger S.M. (2013). 'Between Cheap and Costly Signals: The Evolution of Partly Honest Communication.' *Proceedings of the Royal Society B: Biological Sciences* **280**, 20121878.

Justin Bruner is an associate professor in the Department of Philosophy at the University at Buffalo. His research includes social epistemology, PPE, and the philosophy of biology.

Cite this article: Bruner JP (2024). Assertions: Deterrent or Handicap? A Reply to Graham (2020). *Episteme* 1–11. https://doi.org/10.1017/epi.2023.58