Altruistic Deception

Jonathan Birch
Department of Philosophy, Logic and Scientific Method,
London School of Economics and Political Science,
Houghton Street, London, WC2AE 2AE, UK.

j.birch2@lse.ac.uk
http://personal.lse.ac.uk/birchj1

14 January 2019

Abstract: Altruistic deception (or the telling of "white lies") is common in humans. Does it also exist in non-human animals? On some definitions of deception, altruistic deception is impossible by definition, whereas others make it too easy by counting useful-but-ambiguous information as deceptive. I argue for a definition that makes altruistic deception possible in principle without trivializing it. On my proposal, deception requires the strategic exploitation of a receiver by a sender, where "exploitation" implies that the sender elicits a behaviour in the receiver that is beneficial in a different type of situation and is expressed only because the signal raises the probability, from the receiver's standpoint, of that type of situation. I then offer an example of a real signal that is deceptive in this sense, and yet potentially altruistic (and certainly cooperative): the purr call of the pied babbler. Fledglings associate purr calls with food, and adults exploit this learned association, in the absence of food, to lead fledglings away from predators following an alarm call. I conclude by considering why altruistic deception is apparently so rare in non-human animals.

1. White Lies

Humans tell white lies. Deceiving others for their own good is a normal aspect of human social life. If a student asks whether their essay is the worst in the class, you will probably tell them it isn't, even if, in fact, it is. Concern for the feelings of others sometimes drives us to convey misinformation. White lies are not *always* motivated by concern for others (sometimes we just want to avoid an awkward interaction, for our own benefit) but at least some of them are. When they are motivated by concern for others, they are "altruistic" in the familiar, psychological sense of the word.

Many forms of human altruism have analogues in the non-human world. Compare, for example, Captain Oates sacrificing himself for the sake of his companions with the apparent self-sacrifice of a diseased ant, leaving the colony for the last time to face death alone (Heinze and Walter, 2010). The behaviour of the ants is not altruistic in the *psychological* sense of being motivated by concern for others. But it is *biologically* altruistic, in the sense that it detracts from the viability or fecundity of the actor who performs it, increases the viability or fecundity of another organism (by reducing the risk of disease transmission), and has been maintained by natural selection because of the benefits it confers on others (Birch, 2017).

Deception also has a biological analogue. In the human case, we think of deception as the *intentional* induction by a speaker of a false belief in a listener. Compare this to the "false" alarm calls of the fork-tailed drongo, which scare meerkats away from their prey, allowing the drongo to swoop in and steal the food (Flower, 2011). The drongo may not be intending to induce a false belief in the meerkat. Nevertheless, the false alarm call is *biologically* deceptive: the signal misinforms the meerkat and has, presumably, been maintained by natural selection because of the benefits to the drongo of the behaviour the alarm call induces.

Seeing all this, it is natural to wonder whether their might be cases of *altruistic deception* in the non-human world. A case of altruistic deception would combine the core features of biological altruism and biological deception. A signal would misinform a receiver and would be maintained by natural selection because of this effect. But the selectively relevant benefits would be reaped not by the sender (as in the case of the drongo) but by the receiver (as in the case of white lies). The sender would incur a viability or fecundity cost.[1] Such a behaviour would not require sophisticated theory-of-mind capacities. The misinformative signal would not have to be *intended* by the sender to benefit the receiver by misinforming them; it would simply be maintained by natural selection for this reason. No proximate mechanisms beyond those involved in ordinary cases of signalling would be required. The behaviour would not be a white lie in the human sense, but it would be a biological analogue of a white lie.

Several authors have recently raised the possibility of altruistic deception in something like this sense (Fallis, 2015; Fallis & Lewis, 2017; Artiga & Paternotte, 2018). None, however, has been able to offer a plausible empirical example of altruistic deception in a non-human species. Perhaps this is simply because the phenomenon has not received enough attention from field biologists. However, there is also a conceptual objection to the possibility of altruistic deception: although there is no agreed definition of biological deception, most of the definitions that have been proposed make altruistic deception *impossible by definition*.

The conceptual and empirical debates here are closely entangled. Since biological deception is a term of art in behavioural ecology, we have some flexibility in how we define it. A definition that makes room for the possibility of altruistic deception is desirable only if there are, or plausibly could be, real empirical phenomena that realize that possibility, and only if these phenomena bear enough similarity to cases of non-altruistic deception to make it useful, rather than misleading, to group the altruistic and non-altruistic cases together under the same term. It is therefore a problem for those who want to allow for altruistic deception that they have so far been unable to produce any convincing non-human examples.

[1] If we found a benefit to the receiver without any cost to the sender, we could call this *cooperative* deception but not *altruistic* deception.

My aim in this paper is to weigh into both debates at once. I want to make a plea, based on both conceptual and empirical considerations, for the possibility of altruistic deception. In Section 2, I critically evaluate existing definitions of deception. Definitions in recent literature are more or less split down the middle on the question of whether altruistic deception is impossible by definition, although a slender majority leans towards its impossibility. The reason it tends to be ruled out is that biologists want to characterize deception *without* imputing false content to the signal, and the most natural way to do that is to replace conditions that pertain to the content of the signal with conditions that pertain to the effects on the receiver. In short, the receiver must be harmed. Definitions that don't rule out altruistic deception are diverse, but all face significant problems.

In Section 3 (and readers who just want the positive proposal, and not the evaluation of existing proposals, should skip straight to this), I propose my own definition: a definition that does not rely on imputing false content, but also avoids defining deception in terms of current effects on the receiver. Instead, my definition appeals to an explanatory asymmetry regarding the reasons why the receiver responds in the way it does. Deception, I argue, requires the *strategic exploitation of a receiver by a sender*, where "exploitation" implies that the sender elicits a behaviour in the receiver that is beneficial in a different type of situation and is expressed only because the signal raises the probability, from the receiver's standpoint, of that type of situation.

In Section 4, I turn to the question of whether altruistic deception exists in nature, offering an example of a behaviour that at least potentially meets my proposed conditions. My example is the purr call of the pied babbler (*Turdoides bicolor*). Fledglings associate purr calls with food, and adults exploit this association, in the absence of food, to lead fledglings away from predators following an alarm call (Radford & Ridley, 2006; Raihani & Ridley 2007, 2008). I conclude by considering why altruistic deception, while so familiar in humans, is apparently so rare in non-human animals. I suggest that the special kind of informational asymmetry required for cooperative or altruistic deception (whereby the sender knows more than the receiver not just about the state of the world, but also about how the receiver's interests are best served in that state of the world) requires unusual ecological circumstances.

2. Is Altruistic Deception Impossible by Definition?

2.1 Yes!

2.1.1 Maynard Smith & Harper

Some ways of defining biological deception make altruistic deception impossible by definition. An influential example is the definition of Maynard Smith & Harper:

> Consider a signal that is given in more than one circumstance, but always produces the same response in receivers. Receivers usually benefit from their response, but deception can occur if there is another circumstance in which the same response benefits the sender at the receiver's expense. (Maynard Smith & Harper, 2003, p. 86)

The reason all cooperative deception is ruled out by this definition is the requirement that the receiver is harmed, which I take to imply a reduction in its classical Darwinian fitness (its viability and/or fecundity). Let us call any such condition a *receiver detriment* condition. On standard definitions of biological cooperation, a behaviour cannot be considered cooperative if the recipient of the behaviour is harmed by it (Hamilton, 1964; West et al., 2007; Birch, 2017).

I take it that Maynard Smith & Harper include a receiver detriment condition because, without it, their definition would be far too permissive: it would count as deception any case in which a signal is given in more than one circumstance, benefits the receiver in one circumstance (and may or may not benefit the sender), and benefits the sender in another circumstance (and may or may not benefit the receiver). Such a minimal account would count as deceptive any signalling that is mutually beneficial and involves the use of the same signal in more than one state of the world. Clearly, something must be added to this minimal account.

An intuitive suggestion is to add a *false content* condition: a deceptive signal is one that carries false content, benefits the sender, and would, if true, have benefited the receiver (cf. Searcy & Nowicki, 2005, discussed below). But Maynard Smith & Harper steer clear of this, and I think they are right to do so. I think (in agreement with Skyrms, 2010) that animal signals carry false propositional content only in special circumstances[2], and that we want the notion of biological deception to apply more broadly, whether or not the signals carry false propositional content.[3] Hence the attraction of a receiver detriment condition.

Even without a receiver detriment condition, the *sender benefit* condition of the Maynard Smith & Harper account may be enough to rule out altruistic deception, depending on how payoffs are measured. If payoffs are measured in classical fitness, the sender benefit condition rules out altruism on the part of the sender, because altruism requires that the actor responsible for the social behaviour in question incurs a classical fitness cost (Hamilton, 1964; West et al., 2007; Birch, 2017). If the payoffs are measured in inclusive fitness, altruism is not ruled out—but calculating inclusive fitness payoffs in non-additive interactions

[2] These are, broadly speaking, circumstances in which the population is approximating, but slightly deviating from, a *separating equilibrium* (see Birch, 2014). This account has notable rivals (e.g. Shea et al. 2017), but I suspect that any adequate account of false propositional content in animal signals will vindicate the idea that it arises only in special circumstances. For example, Shea and colleagues' account only allows for propositional content when the population is at an equilibrium (see Section 2.1.4).

[3] We might also want the notion of biological deception to apply to states or processes that are not signals, because they don't play the right kind of mediating role between a sender and a receiver (Artiga and Paternotte 2018). Consider, for example, fixed camouflage that the sender is unable to alter flexibly. I don't discuss examples of this type here; my topic is what it takes a for a signal to be biologically deceptive.

such as signalling games, without double-counting, presents non-trivial technical challenges (as discussed by Birch, 2013, 2016, 2017; Okasha and Martens, 2016a, b).

2.1.2 Skyrms

Skyrms (2010) starts with a definition of *misinformation* that is intentionally broader than the concept of deception:

> If receipt of a signal moves probabilities of states it contains information about the state. If it moves the probability of a state in the wrong direction—either by diminishing the probability of the state in which it is sent, or raising the probability of a state other than the one in which it is sent—then it is misleading information, or *misinformation*. (Skyrms, 2010, p. 80).

The probabilities here are objective probabilities *from the receiver's standpoint*: they do not reflect the receiver's degrees of belief (it need not have any degrees of belief) but rather the objective probabilities of states of the world, conditional on the signals the receiver has received. Whether a signal *raises* or *lowers* the probability of a state of the world, from the receiver's standpoint, is evaluated by comparing the objective probability of the state *prior* to receiving the signal with its objective probability after conditionalizing on the signal.

Misinformation in Skyrms's sense encompasses signals that raise the probability of the true state of the world, but at the same time also raise the probability of a non-actual state (by, for example, ruling out all but two states). In other words, it encompasses cases in which the sender provides useful-but-ambiguous information to the receiver about the state of the world. I take it we do not want to classify *all* such signals as deceptive. Sometimes, useful-but-ambiguous information arises due to an information bottleneck in which there are too few signals available to allocate one signal to every state of the world. In these cases, the sender may be doing the best it can to help the receiver with the imperfect tools it has available.

To avoid classifying such cases as deception, Skyrms adds three further conditions: systematicity, sender benefit and receiver detriment. He writes:

> If misinformation is sent systematically and benefits the sender at the expense of the receiver, we will not shrink from following the biological literature in calling it *deception*. (Skyrms, 2010, p. 80)

Skyrms suggests in a footnote that "not much hangs" on whether the receiver detriment condition is included. He suggests we could use the terms *weak* and *strong* deception, using the latter to mark those cases involving receiver detriment. The category of weak deception would then be very broad, encompassing all cases of useful-but-ambiguous information that systematically generates a benefit to the sender. Weak deception could be cooperative, but could it be altruistic? As noted above, this depends on how payoffs are measured. If payoffs are measured in classical fitness, the sender benefit condition alone rules out altruistic deception.

2.1.3 McWhirter

McWhirter (2016) argues that we should replace the *misinformation* component of Skyrms's account with *misuse*. For McWhirter, a signal is misused by a type of sender if, conditional on it being sent by that type of sender, it shifts probabilities differently to the way it shifts probabilities when sent by the average sender.

Because McWhirter endorses Skyrms's sender benefit and receiver detriment conditions, his account rules out cooperative or altruistic deception, while allowing for a much broader category of cooperative or altruistic misuse. However, this broader category will include cases in which a sender misuses a signal only because it has found a way of conveying *more* information to the receiver (e.g. perhaps it has overcome an information bottleneck to which the average sender is subject). Misuse alone is too broad to mark out a category of deception, yet the narrower category of receiver-harming misuse cannot be cooperative or altruistic.

2.1.4 Shea et al.

Shea et al. (2017), unusually, define deception in terms of sender benefit, receiver detriment *and* false propositional content:

> We understand deception to occur when a message with a false content is sent and the receiver is induced to behave in a way that benefits the sender and harms the receiver. (Shea et al., 2017, p. 18)

Because Shea et al. require false propositional content for deception, their account has the drawback that it only allows for deception in a narrowly circumscribed set of cases. Given their account of propositional content, determinate propositional content exists only when the population is at an evolutionary equilibrium. There is no scope here for away-from-equilibrium deception.

Leaving this concern aside, it is puzzling that Shea et al. include a receiver detriment condition in addition to a false content condition, since the usual motivation for a receiver detriment condition is narrow down the category of deception *without* requiring the attribution of false content. This leads to the question: if we were to strike out the receiver detriment and sender benefit conditions, and take false content alone (in the sense of Shea et al.) to be constitutive of deception, might the Shea et al. definition be compatible with altruistic or cooperative deception after all?

To find out, we first need to understand the basis on which Shea et al. attribute false content to a signal. They offer a subtle way of attributing content based on the payoffs attained by sender and receiver. We first consider the "baseline" payoffs for sender and receiver, which are the payoffs attained if the receiver performs the behaviour that maximizes expected payoff without conditionalizing on any signal. We can then ask, for any particular signal *S*, whether its transmission in a given state of the world *X* induces a behaviour in the receiver that leads to an "above-baseline" payoff for both parties. If it does, we can say that *X* is part of the content of *S*. The overall content of *S* is the disjunction of states in which it leads to above-baseline payoffs. If *S* is then sent in a state of the world in which it fails to lead to above-baseline payoffs, we can say it carries false content. For example, an alarm call that *would* lead to above-baseline payoffs in the presence of a leopard, but which does not lead to above-baseline payoffs in the absence of a leopard, can be said in the latter case to be falsely representing the presence of a leopard.

On close inspection, this procedure for attributing false content, combined with a definition of deception that counts false content alone as sufficient for deception, allows for cooperative or altruistic deception when both of the following conditions obtain:

i. Sending *S* in *X* benefits the receiver (in the sense of raising its payoff in comparison with what it would otherwise have been) but does not lead to above-baseline payoffs for both sender and receiver.
ii. There is some other state (or disjunction of states) of the world, *X*′, such that sending *S* in *X*′ does lead to above-baseline payoffs for both sender and receiver.

In these circumstances, the signal will falsely represent *X*′, but it will benefit the receiver by doing so. I suggest, therefore, that Shea et al. can be more open to the possibility of cooperative or altruistic deception than their official definition of deception suggests.

2.2 No!

2.2.1 Searcy & Nowicki

Accounts of deception that omit a receiver detriment condition have drawbacks of their own. A notable example is the definition of Searcy & Nowicki (2005, p. 5), who define deception as occurring when:

1. A receiver registers something *Y* from a signaler;
2. The receiver responds in such a way that
   a. Benefits the sender and
   b. Is appropriate if *Y* means *X*; and
3. It is not true that *X* is the case.

This is an example of account that defines deception in terms of imputed false content (*X* denoting the proposition in question), without providing clear conditions for attributing content to signals, and can be criticized on these grounds. If we want a concept of deception broad enough to cover cases where no determinate propositional content can be attributed to the signal, we must define deception in other terms (Skyrms, 2010, p. 76).

It can also be criticized for its reliance on “appropriateness”. In saying that the receiver’s response is “appropriate if *Y* means *X*”, Searcy and Nowicki avoid explicitly requiring that *Y* means *X*, which may initially seem to be a virtue of the account. But, taken literally, the condition is far too permissive. Fleeing would be an appropriate response to an alarm call if that alarm call meant “a 4x4 full of poachers is coming”, but the call is not thereby deceptive in all cases in which the sender benefits and no 4x4 full of poachers is nearby. What matters for deception is the signal's actual content, not whether there is a possible content that would rationalize the receiver's response.

2.2.2 Artiga & Paternotte

Artiga & Paternotte (2018) define a deceptive state (or signal) as a state (or signal) with the *function* of causing a misinformative state, where function is defined along the lines of an etiological theory of function: a signal that has the function of causing a misinformative state is one that has, in recent history, been selected-for because of the fitness consequences of the misinformative state it induces in the receiver.[4] Since these consequences may or may not include receiver detriment, the account is compatible with altruistic or mutually beneficial deception, as Artiga & Paternotte emphasize.

The implications of this proposal depend on what is meant by "a misinformative state". Artiga & Paternotte do not endorse any particular account. At face value, talk of misinformative states may suggest that the receiver possesses a belief-like representation of the world that can be evaluated as true or false, restricting the scope of the account to animals with such representations. However, the requirement might be read more minimally as requiring merely that there is some internal state of the receiver, INT, such that $P(X \mid \text{INT}) < P(X)$, where X is the actual state of the world, or $P(X' \mid \text{INT}) > P(X')$, where X' is some non-actual state of the world. This could be a perceptual, neural or even biochemical state that need not be belief-like. A deceptive signal is one with the function of causing the receiver to possess such a state. Artiga & Paternotte (2018, p. 592) seem to lean towards this more minimal reading. To be clear, Artiga and Paternotte do not explicitly endorse this or any other account of misinformation; but it is an example of a possible account of misinformation with which their basic proposal could be combined.

The drawback is that the resulting concept of deception is too permissive: it makes altruistic or mutually beneficial deception too easy, by classing as deceptive exactly the sort of mutually beneficial, useful-but-ambiguous signalling in cases of information bottlenecks that needs to be ruled out. Suppose that, due to an information bottleneck, the same signal is sent in two states, X and X'. Whichever of these states obtains, the internal state of the receiver induced by the signal, INT, will be misinformative in Skyrms's minimal sense, since, although it raises the probability of the actual state of the world, it also raises the probability of a non-actual state. If we add that the signal also has the function of causing INT, all of Artiga & Paternotte's conditions for deception are met.

2.2.3 Fallis & Lewis

Fallis & Lewis (2017) argue that all of the preceding accounts fail to appreciate the significance of *accuracy* to deception. A deceptive signal, they propose, is one that benefits the sender by shifting the probabilities of states of the world, from the standpoint of the

[4] On etiological theories of function, see Wright (1976); Millikan (1984); Neander (1991); Godfrey-Smith (1994).

receiver, in a way that reduces their accuracy. Generally speaking, signals that shift probability away from the true state of the world, and towards a non-actual state, will reduce accuracy, but the precise relation between probabilities and accuracy is subtle and depends on the formal accuracy measure one uses. This ties the literature on biological deception to formal epistemology, in which one major on-going debate concerns the best formal measure of accuracy (e.g. Joyce, 1998; Pettigrew, 2016).

The connection to formal epistemology might be seen as an advantage or a drawback, depending on how hopeful one is that epistemologists will converge on an agreed measure of accuracy. Their preferred measures are "proper scoring rules", such as the Brier rule, the log rule and the spherical rule, which were originally developed by statisticians for evaluating the accuracy of forecasts (e.g. weather forecasts). These rules are "proper" in the sense that they are uniquely optimized by probability distributions that line up with objective chances. But these scoring rules often disagree with each other in realistic cases, and epistemologists cannot agree on the correct scoring rule, nor can they agree on what it is that makes a proper scoring rule (as opposed to some other scoring rule) the right way to measure accuracy. Many of the key players in the debate insist that the justification for the one true scoring rule, whatever it is, must be a *non-pragmatic* justification, or else arguments for Bayesian principles based on scoring rules are no better than traditional pragmatic arguments. Such a justification remains elusive, and the possibility of such a justification remains contentious.

In the absence of a single agreed measure, one can still find (as Fallis & Lewis do) cases in which a signal reduces accuracy by the lights of any reasonable measure. But it is unclear how to handle cases in which a signal reduces accuracy according to one rule but not according to another. Perhaps more seriously, given that our aim here is develop a concept of *biological* deception, it is unclear how biologists are to go about applying an accuracy measure in the field. If the aim here is a concept of deception justified by its utility in guiding inquiry in behavioural ecology, it seems advisable *not* to require an assessment of the receiver's probability distribution using a proper scoring rule.

Fallis & Lewis's proposal also faces notable problem cases. Skyrms's paradigm case of biological deception is mimicry in the firefly genus Photuris. Photuris will mimic the mating signal of a female firefly of the genus Photinus, use it to lure in Photinus males, and eat them. Skyrms asks, rhetorically, "I would say that this qualifies as deception, wouldn't you?" (Skyrms, 2010, p. 75). Fallis & Lewis have to answer that it would not (Fallis & Lewis, 2017, pp. 10-11). Photuris's mating signal rules out one state of the world from the receiver's standpoint (a state in which there is no Photinus and no Photuris nearby) and raises the probability of both a nearby Photinus female and a nearby Photuris. Probabilities, conditional on the signal, are more, not less, accurate than the unconditional probabilities, according to any of the usual measures of accuracy, because one possible state of the world has been correctly ruled out. This is not a decisive consideration, but it does indicate a significant misalignment between Fallis & Lewis's proposal and biological practice, since biologists take cases of this general type to be among the core examples of biological deception.

3. A Synthesis: Strategic Exploitation of Receivers by Senders

All else being equal, I side with the "No!" camp: I think we should prefer a definition of biological deception that does not rule out cooperative or altruistic deception *a priori*. To explain why, I need to say something about the criteria for evaluating a definition of a theoretical term. Compatibility with our pre-theoretical intuitions, or otherwise, is of little relevance: there should be some rationale for co-opting an ordinary language term, but we should not be concerned if its usage as a term of art in behavioural ecology departs significantly from its ordinary usage (cf. altruism, spite, cooperation). So, the fact that there can *intuitively* be cases of altruistic deception is not a strong objection to definitions that rule it out.[5] It is a more serious problem if the definition fails to encompass cases, such as the Photuris/Photinus case, that biologists regard as core examples of biological deception. But alignment with our ordinary, pre-theoretical intuitions about deception is of no great significance.

What *does* matter, however, is that a theoretical term is able to *guide inquiry in productive ways*.[6] Maynard Smith & Harper, Skyrms and McWhirter all start with a basic platform that is too broad to capture biological deception, and then add sender benefit and receiver detriment conditions as a way of narrowing the extension of the concept. An account of this type is guiding inquiry in a particular way: it says, to identify biological deception, first look for the minimal feature of the broader category (signals sent in more than one circumstance, misinformative signals, or misused signals), and then look for sender benefit and receiver detriment. Such an account guides us away from looking for cases that substantially resemble non-cooperative cases of biological deception but in which the receiver benefits from being deceived, or in which the sender is harmed by the deception. If there are such cases (on this, see Section 4), definitions of this type will guide researchers away from finding them. If any such behaviours are found, researchers will be led to describe these deception-like behaviours in other terms, obscuring their similarity to non-cooperative cases of deception and obstructing a productive line of inquiry.

This is not a decisive consideration, but it is a consideration that matters. All else being equal, we should prefer a definition of a theoretical term that does *not* obstruct a productive line of inquiry. Of course, all else may not be equal. But this consideration gives us reason to look for other ways of defining deception that allow us to talk about cooperative and altruistic deception as genuine empirical possibilities (a point also made by Artiga & Paternotte, 2018 and Fallis & Lewis, 2017).

[5] On this point I disagree with Artiga & Paternotte (2018) and Fallis & Lewis (2017).

[6] Compare, in this context, recent arguments for using a concept of "reciprocal causation" in preference to Mayr's (1961) "proximate-ultimate" distinction. These arguments share my pragmatic evaluation criterion: reciprocal causation is defended on the grounds that it steers inquiry in productive ways (Laland et al., 2011).

We have seen, however, that existing accounts of deception that aim to make room for the possibility of altruistic deception face problems of their own. They require us to impute false content to signals on unclear grounds (Searcy & Nowicki), or they fail to rule out cases in which the sender informs the receiver as best it can with imperfect tools (Artiga & Paternotte), or they fail to rule *in* cases in which a sender exploits a receiver while also raising the probability of the true state of the world (Fallis & Lewis).

Is there any way to synthesize the insights of all these accounts to create a definition that captures the core features of biological deception *without* ruling out altruistic deception *a priori*? Such an account would have to capture what is deceptive about Photuris luring Photinus, while also capturing what is non-deceptive about a sender conveying partial information only because its tools are limited. Moreover, it would have to do this without introducing sender benefit and receiver detriment conditions, and without requiring false propositional content.

Here is a proposal. What cases of deception have in common, and what sets them apart from cases of non-deceptive signalling, is the *strategic exploitation* of receivers by senders. Recognizing the importance of exploitation pulls us towards introducing a receiver detriment or false content condition; but there is another, more minimal way to capture the relevant kind of exploitation. In a case of deception, a signal S, sent in a state of the world X, induces a behaviour B in a receiver.[7] The functional explanation for why the receiver responds to the signal by expressing B is that B is beneficial in some other state of the world, X', and S raises the probability of X' from the receiver's standpoint (that is, the signal is such that $P(X'|S) > P(X')$). *The sender strategically exploits an adaptive disposition in the receiver by raising the probability, from the receiver's standpoint, of a non-actual state of the world.*

Strategic exploitation in this sense does not require that the receiver is *harmed* by expressing B in X. The core feature of the pattern is an *explanatory asymmetry*: the functional explanation for why B is expressed in X is that it benefits receivers in another state of the world, X'.[8] The possibility of B also benefiting the receiver in X is not logically ruled out: what matters is that any benefit to the receiver in X is not part of the *explanation* for why the receiver expresses B in X.

The relevant sort of "explanation" here is functional (or teleological) explanation in the sense of Wright (1976) and Neander (1991): to say that the benefit to the receiver of expressing B in X' explains its expression in response to S is to say that the *past consequences* of expressing B as a response to S in X' have fed back into a process of natural selection or learning in such a way as to be partly responsible for the receiver's current performance of B. It is to say, in

[7] Linking deception to action in this way implies that a signal cannot be deceptive if the receiver fails to respond. Thus "deceive" here is interpreted (as in Shea et al., 2017) as a "success verb", a verb that implies the successful completion of the process it describes.

[8] This part of the proposal is reminiscent of (and probably inspired by) Fodor's (1987) "asymmetric dependence" theory of content, developed in a rather different context to solve a rather different problem.

Wright's memorable phrase, that there is a *consequence-etiology* behind the expression of B in X'. There is an explanatory asymmetry between the two uses of the signal if the same cannot be said of X: that is, if there has been no feedback from expressing B in X to a process of natural selection or learning.

This asymmetry is related to, but not the same as, Godfrey-Smith (2011) and Shea and colleagues' (2017) distinction between maintaining (or stabilizing) and non-maintaining (or destabilizing) uses of a signal. The explanatory asymmetry to which I am appealing concerns the (evolutionary or learning) history of the signal, not its current effects. On my account, a signal may still be deceptive, and exploit the adaptive dispositions of the receiver, even if it benefits the receiver on the occasion in question. This is what happens in cases of altruistic or cooperative deception, for which the account is intended to allow.

It is not even necessary, I contend, that the sender benefits. What is required is that the signal is sent in *X strategically*, which is to say simply that it is sent as part of the sender's strategy (cf. Skyrms's "systematicity" requirement). The need to capture the strategic element of deception is what motivates a sender benefit requirement, but the requirement is too heavy-handed: a strategy need not benefit the actor who implements it. A strategy is any pattern of behaviour that can be transmitted, by some inheritance system or other, down the generations, and that can be characterized with conditionals of the form "If in context C, perform behaviour B" (Maynard Smith, 1982; Birch, 2017). Strategies, not individual actions, are the main explanatory targets of behavioural ecology. What matters for our purposes is that some signals are sent *as parts of strategies* and some are not. The ones that are not are mere accidents or one-offs: there is no transmissible pattern to them, and they are of no particular interest to behavioural ecologists. The notion of biological deception is intended to pick out a particular aspect of a strategy.

To rule out cases in which a behaviour is induced in the receiver as a mere by-product of a strategy with a completely unrelated function, we should require that strategy has been maintained by selection at least in part because of the payoffs conferred by the receiver's performance of B in X. However, we need not require that the relevant payoffs are conferred *on the sender* as opposed to the receiver. Given a background theory of function that ties function to recent selection history, this amounts to saying that eliciting the performance of B in X must be part of the strategy's function (and this is a point of agreement with Artiga & Paternotte, 2018). But since such theories of function are not universally accepted, I prefer to formulate the relevant condition without using the word "function".

The *sender strategy* and *receiver exploitation* elements of deception are combined in the following account:

A signal S, sent in a state of the world X, is *biologically deceptive* if and only if:

> *Receiver exploitation conditions:*
>
> a) Sending S in X elicits some behaviour B in the receiver.

b) *S* elicits *B* in *X* not because *B* benefits receivers in *X*, but because (i) *B* benefits the receiver in some other state of the world, X', and (ii) $P(X'|S) > P(X')$.

Sender strategy conditions:

c) *S* is sent in *X* as part of a strategy.

d) The sender's strategy has been maintained by selection at least in part because of the payoffs conferred by receivers' performance of *B* in *X*.

Does this account have the features we wanted it to have? The account captures what is deceptive about Photuris: a signal, sent as part of a strategy, exploits an adaptive disposition in the receiver (mating behaviour) by raising the probability, from the receiver's standpoint, of a non-actual situation (the presence of a potential mate). The account also captures what is non-deceptive about cases in which the sender is merely constrained by an information bottleneck. In these cases, the explanation for the receiver's behaviour is that it is advantageous in both of the states indicated by the signal. The signal is not deceptive because the crucial explanatory asymmetry is not present.

Because the account avoids a sender benefit or receiver detriment condition, it makes conceptual room for the possibility of cooperative or altruistic deception. We can now ask: are there empirical examples that actually occupy that room? In other words, are there empirical cases in which a sender strategically exploits an adaptive disposition in the receiver by raising, from the receiver's standpoint, the probability of a non-actual situation… to the receiver's *benefit*?

4. Cooperative Deception in Pied Babblers

The southern pied babbler (*Turdoides bicolor*) is a bird species found in the dry savannah of southern Africa. Pied babblers are cooperative breeders, living in family groups of 3-10 adults and their offspring. Amanda Ridley's Pied Babbler Research Project has studied a fully habituated population of these birds in the Kalahari Desert since 2003, providing numerous insights into their behaviour and one potential example of cooperative, and perhaps altruistic, deception.

Adult pied babblers give a "purr call" when feeding nestlings and fledglings (Radford & Ridley, 2006). The nestlings, by the age of 14 days, learn to associate this call with food. This learned association causes the nestlings to express begging behaviour. Support for the claim that nestlings associate the purr calls with food comes from evidence that their responses to the purr calls are hunger-dependent: they are more likely to beg when they have not recently fed (Raihani & Ridley, 2007). Support for the claim that the association is learned, rather than innate, comes from evidence that exposing the nestlings to playbacks of purr calls every time an adult brings food will increase the speed with which they develop the association (Raihani & Ridley, 2008). Raihani and Ridley argue that this may be interpreted as a basic form of teaching, since the most plausible function of the purr calls made during feeding is to facilitate the learning of an association between purr calls and food, which is then useful during the fledgling phase. Teaching or not, it is a form of active classical conditioning of nestling behaviour by adults.

So far, the purr call sounds like a normal case of honest signalling. But in the days immediately after the offspring have fledged (i.e. left the nest), and especially on the day of fledging itself, adults use these same purr calls in other contexts, in the absence of food (Raihani & Ridley, 2007). They give the same calls when moving between foraging sites, inducing the fledglings to follow (Radford & Ridley, 2006). In these contexts, the call might still be interpreted as meaning "food here!", albeit in a broader sense than previously. More strikingly, however, the adults also use the purr call in the minutes immediately after an alarm call (Raihani & Ridley, 2007). These post-alarm purr calls induce the fledglings to move towards the adult, away from the nest and away from where the predator has been seen.

The use, in the absence of food, of a call that the fledglings associate with food, in order to induce them to move away from a predator, invites an intuitive description in terms of deception. It is intuitive to say that the adult, by calling "food here!" to the fledgling, tells a white lie that elicits movement and thereby protects the fledgling from predation. Of course, as I emphasized in the preceding sections, we should be cautious about imputing false content to the signal. We can, however, ask whether it meets the conditions for deception in my proposed account.

Are the receiver exploitation conditions satisfied? Yes. The fledgling's response to the purr call—approaching the adult and begging—is explained not by any prior learning of, or selection for, this behaviour in the presence of a predator (the actual state of the world). The fledgling will, in time, learn to respond to alarm calls as adults do, but that is not what explains its response to the purr call. The behaviour is explained by prior learning of an association between this behaviour and positive payoffs in another, non-actual state of the world: one in which the adult has food. The signal induces that behaviour because it raises the probability, from the fledgling's standpoint, of the adult possessing food.

Are the sender strategy conditions satisfied? Yes. The active conditioning of the young by adults, so that they associate purr calls with food, followed by the use of that same call in non-feeding contexts to elicit movement, is a systematic pattern of observed behaviour. It may be transmitted down the generations genetically or by learning, but the precise mechanism is not important to its status as part of a strategy. Moreover, the strategy has an adaptive rationale: by helping the fledglings avoid predators, the adult, which will normally be related to the fledging if not its parent promotes the viability of a relative. If the sender's strategy is genetically inherited, the benefits will fall differentially on other bearers of the genes in question; if the strategy is learned, the benefits will fall differentially on recipients likely to learn that same strategy in later life. Either way, we can reasonably hypothesize that the "white lie" behaviour has been maintained by selection because of the benefits conferred by the successful evasion of predators by the fledglings.

In sum, there is a good case for regarding the use of purr calls in non-feeding contexts as biologically deceptive. But is the deception *altruistic*? It is certainly *cooperative*, if we grant that the deceived fledglings benefit and that the sender's strategy has been maintained because of this benefit. But is a harm imposed on the sender, measured in the currency of lifetime reproductive success? We should focus here on the cases in which the adult is not the parent of the fledgling but another group member (benefits conferred on offspring by their own parents are not normally considered biologically altruistic, because they enhance the parent's lifetime reproductive success). Even in these cases, there is insufficient evidence to conclude with any certainty that the behaviour is altruistic, because we can't be sure that there is a lifetime fitness cost.

This is, however, a serious empirical possibility. Alarm calls are good candidates for biological altruism, because the sender makes an audible call that may draw a predator's attention to itself. Of course, if it does draw the attention of a predator, the benefit to the fledglings will presumably be cancelled, so the success of the strategy relies on this being unlikely. Yet even if the call draws the attention of a predator only rarely, it is likely to make the behaviour costly to the sender's classical fitness overall, since there will be some reduction in the sender's viability, and no apparent benefit to the sender that might outweigh this cost.

In short, we have here a plausible example of *cooperative* deception that may, or may not, also qualify as a case of *altruistic* deception in some circumstances.

5. Why So Rare?

I have only come across one example of cooperative or altruistic deception that I find persuasive. One is enough to show that the category is not empty and deserves further exploration, but it is not many. Why has behavioural ecology not uncovered more examples? One reason may be that definitions of deception have tended to discourage looking for such phenomena—a manifestation of one of the potential drawbacks to such definitions I highlighted in Section 3.

It may also be that there are more examples in the literature, but that they are hard to find because the term "deception" is not used and the behaviour is instead described in a way that conceals, rather than makes perspicuous, the strategic exploitation of receivers by senders. That is certainly true of Raihani & Ridley's example: they do not use the term "deception" in print, despite the intuitively deception-like character of the behaviour. This illustrates another potential drawback of defining deception overly narrowly.

However, another plausible reason is that cases of cooperative or altruistic deception *really are* rare in nature, because cooperative or altruistic deception arises only in unusual conditions. The example of the pied babbler suggests a general place to look for examples: adult-juvenile interactions in which the juvenile is too inexperienced to respond optimally to every signal used by the adults. If there are alternative signals the adult can send that will induce a somewhat appropriate response by exploiting existing learned associations, the juvenile may be deceived and yet benefit from the deception. Yet in addition to pointing towards a potentially fruitful line of empirical investigation, this example also hints at the probable rarity of the conditions for cooperative or altruistic deception.

The phenomenon requires a special kind of informational asymmetry between sender and receiver. For any kind of signalling to arise, the sender must know something about the state of the world that the receiver does not know. For cooperative deception to arise, the sender must additionally know more than the receiver about the *best response* to the current state of the world *given the receiver's interests*. More precisely, there must be a history of learning and/or evolution behind the sender's behaviour in *X* that explains its use of the deceptive signal, and the effects of receiver behaviour must have fed back into that process. Yet there must be no corresponding history of learning and/or evolution behind the current receiver's behaviour in *X*: for the signal to count as deceptive, the learning or evolution that has shaped the receiver's behaviour in *X* must have occurred in a different environmental state. So, we need special conditions: *a sender whose strategy has been shaped by past interactions in this state of the world, and a receiver whose strategy has not*. In addition, there must be enough alignment of interest between sender and receiver to support cooperation or altruism. We can imagine how these conditions may arise in the context of adult-juvenile interactions, where, in species capable of learning, there is inevitably a large asymmetry of experience. But it is hard to imagine how they could arise outside of that context.

Acknowledgements

This article is a contribution to a special issue on Content in Signalling Systems. I thank Marc Artiga, Nichola Raihani and two anonymous reviewers for their comments. This work was supported by a Philip Leverhulme Prize from the Leverhulme Trust.

References

Artiga, M., & Paternotte, C. (2018). Deception: a functional account. *Philosophical Studies*, *175*, 579-600.

Birch, J. (2014). Propositional content in signalling systems. *Philosophical Studies*, *171*, 493-512.

Birch, J. (2013). Review of S. Okasha and K. Binmore (Eds), Evolution and Rationality: Decisions, Cooperation, and Strategic Behaviour. *British Journal for the Philosophy of Science*, *64*, 669-673.

Birch, J. (2016). Hamilton’s two conceptions of social fitness. *Philosophy of Science*, *83*, 848-860.

Birch, J. (2017). *The Philosophy of Social Evolution*. Oxford: Oxford University Press.

Fallis, D. (2015). Skyrms on the possibility of universal deception. *Philosophical Studies*, *172*, 375-397.

Fallis, D., & Lewis, P. J. (2017). Toward a formal analysis of deceptive signaling. *Synthese*. Advance online publication. doi:10.1007/s11229-017-1536-3

Flower, T. (2011). Fork-tailed drongos use deceptive mimicked alarm calls to steal food. *Proceedings of the Royal Society of London B: Biological Sciences*, *278*, 1548-1555.

Fodor, J. A. (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Godfrey-Smith, P. (1994). A modern history theory of functions. *Noûs*, *28*, 344-362.

Godfrey-Smith, P. (2011). Review of Signals: Evolution, Learning, and Information, by Brian Skyrms. *Mind*, *120*, 1288-1297.

Hamilton, W. D. (1964). The genetical evolution of social behaviour I and II. *Journal of Theoretical Biology*, *7*, 1-52.

Heinze, J. and Walter, B. (2010). Moribund ants leave their nests to die in social isolation. *Current Biology*, *20*, 249-252.

Joyce, J. M. (1998). A non-pragmatic vindication of probabilism. *Philosophy of Science*, *65*, 575-603.

Laland, K. N., Sterelny, K., Odling-Smee, J., Hoppitt, W., & Uller, T. (2011). Cause and effect in biology revisited: Is Mayr’s proximate-ultimate dichotomy still useful? *Science*, *334*, 1512-1516.

McWhirter, G. (2016). Behavioural deception and formal models of communication. *British Journal for the Philosophy of Science*, *67*, 757-780.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

Maynard Smith, J., & Harper, D. (2003) *Animal Signals*. Oxford: Oxford University Press.

Mayr, E. (1961). Cause and effect in biology. *Science*, *134*, 1501-1506.

Millikan, R. G. (1984). *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Neander, K. (1991). The teleological notion of ‘function’. *Australasian Journal of Philosophy*, *69*, 454-468.

Okasha, S., & Martens, J. (2016a). Hamilton’s rule, inclusive fitness maximization, and the goal of individual behaviour in symmetric two-player games. *Journal of Evolutionary Biology*, 29, 473-482.

Okasha, S., & Martens, J. (2016b). The causal meaning of Hamilton’s rule. *Royal Society Open Science*, *3*, 160037. doi:10.1098/rsos.160037

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Radford, A. N. & Ridley, A. R. (2006). Recruitment calling: a novel form of extended parental care in an altricial species. *Current Biology*, *16*, 1700-1704.

Raihani, N. J., & Ridley, A. R. (2007). Adult vocalizations during provisioning: offspring response and postfledging benefits in wild pied babblers. *Animal Behaviour*, *74*, 1303-1309.

Raihani, N. J., & Ridley, A. R. (2008). Experimental evidence for teaching in wild pied babblers. *Animal Behaviour*, *75*, 3-8.

Searcy, W. A., & Nowicki, S. (2005) *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton, NJ: Princeton University Press.

Shea, N., Godfrey-Smith, P., & Cao, R. (2017). Content in simple signalling systems. *British Journal for the Philosophy of Science*. Advance online publication. doi:10.1093/bjps/axw036

Skyrms, B. (2010) *Signals: Evolution, Learning, and Information*. Oxford: Oxford University Press.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, *20*, 415-432.

Wright, L. (1976). Functions. *Philosophical Review*, *82*, 139-168.