Received: 16 February 2024 | Accepted: 13 May 2024

DOI: 10.1111/rati.12417

SPECIAL ISSUE ARTICLE

WILEY

The property of goal-directedness: Lessons from the dispositions debate

Matthew Tugby

Department of Philosophy, Durham University, Durham, UK

Correspondence
Matthew Tugby, Department of Philosophy, Durham University, 50 Old Elvet, Durham DH1 3HN, UK.
Email: matthew.tugby@durham.ac.uk

Abstract

The system-property or 'cybernetic' theory of goals and goal-directedness became popular in the twentieth century. It is a theory that has reductionist and behaviourist roots. There are reasons to think that the system-property theory needs to be formulated in terms of counterfactuals. However, it proves to be difficult to formulate a counterfactual analysis of goal-directedness that is counterexample-free, non-circular, and non-trivial. These difficulties closely mirror those facing reductionists about dispositions, though the parallels between the two debates have been overlooked in the literature. After outlining those parallels, the paper considers what goal theorists might learn from the dispositions debate. In particular, the paper discusses the need for a realist, non-reductionist account of goal-directedness, and explores the idea that properties of goal-directedness are themselves dispositions or 'powers' of a certain sort.

KEYWORDS
counterfactuals, dispositions, goal-directedness, goals, powers, reduction

1 | INTRODUCTION

In twentieth-century philosophy of science, the concepts of goals and goal-directedness started to attract a lot of interest from the 1940s onwards. This was partly because of the rise of cybernetic systems in industry

and (regrettably) warfare, such as heat-seeking missiles, speed regulation governors, and thermostats (to name a few). However, teleological concepts were already familiar in biology, where it is natural to describe many organisms and organic subsystems as being goal-directed, as when we say that a falcon's behaviour is directed towards the capture of prey, or that the goal of a functioning kidney is to help moderate blood's water content. In the philosophy of biology, such talk is often grounded in historical evolutionary facts about natural selection: the goal of a heart is to beat because it is by virtue of their beating that hearts were selected and replicated in the first place. But due to the rise of cybernetic systems, it also became natural to ascribe goals and goal-directed behaviour in many non-biological cases. For example, when a heat-seeking missile relentlessly pursues its target, and overcomes obstacles in the process, it displays complex behaviour that is not unlike that of a falcon pursuing its prey. This observation leads to an important question. If it is legitimate to ascribe goal-directed behaviour to many biological and non-biological systems, what is it that such systems have in common? What is it about these diverse systems that makes talk of goal-directedness appropriate in each case? If we could start to answer this question, we would take an important step towards providing a general metaphysical analysis of teleological phenomena that is applicable in different branches of science. Moreover, if we can say what it is that certain biological and non-biological systems have in common, it will be easier to see what their differences are.

In the middle part of the twentieth century, the so-called 'system-property' or 'cybernetic' theorists, many of whom were philosophers of biology, tried to answer the questions above in naturalistic terms. Amongst the early pioneers were Rosenblueth et al. (1943) and Braithwaite (1953), who offered empirical analyses of goal-directedness that were grounded in the patterns of behaviour that goal-directed systems exhibit. Other versions of this approach were also developed, which added various mechanistic or modal constraints on what could count as a goal-directed system (e.g., Nagel, 1961; Sommerhoff, 1950). What all these theories have in common is that they are broadly reductionist: they aim to show that whilst goal-directedness does indeed occur in biology and elsewhere, it can ultimately be explained in a non-teleological way in terms of mechanistic behaviour of a certain sort. According to this approach, what distinguishes goal-directed mechanisms from non-goal-directed mechanisms is merely the level of complexity and adaptability of the former's behaviour. In Nagel's terminology, goal-directed systems are 'directively organized' (Nagel, 1961, p. 415). This naturalistic approach to goal-directedness is sometimes known as the 'cybernetic' theory, whilst Nagel refers to it as the 'system-property' theory.[1] In what follows I shall use Nagel's 'system-property' terminology, because the word 'cybernetic' misleadingly suggests that the theory only concerns artificial, cybernetic machines. As we have already seen, the theory is meant to provide a general account of goal-directed behaviour that is applicable in both biological and non-biological cases, and thereby provide a general analysis of teleological explanation that covers different branches of science.

System-property theories of goal-directedness have several attractive features and, in many cases, they have provided a level of formal and scientific rigour that, arguably, has not been matched in this area. Nonetheless, these approaches face some serious difficulties. One of the most pressing problems concerns how system-property theorists can accommodate and explain goal-directedness in cases where the relevant system *fails* to achieve its alleged goal (e.g., Ehring, 1984a; Nissen, 1997; Scheffler, 1963; Tugby, 2024, sect. 3). Let us call this the *problem of failure*. As we shall see, the most obvious way of overcoming the problem of failure is to express the system-property theory in counterfactual terms. The counterfactual proposal is that even if a directively organised system does not *actually* achieve its goal, it might still count as being goal-directed in virtue of certain complex behaviours that it *would* exhibit if certain possible (but non-actual) circumstances were to obtain. In the following three sections, we shall examine the prospects for this counterfactual version of the system-property theory. Unfortunately, we shall see that the most obvious counterfactual formulations either face counterexamples or become vacuous. I am not the first person to point out some of these problems. However, what has gone

[1]To add further terminological variation, Nissen (1981) calls Nagel's theory the 'self-regulation' analysis.

unnoticed is that there are close parallels between the attempt to reductively analyse goal-directedness using counterfactuals and later attempts by philosophers to reductively analyse *disposition* predicates. Such parallels have been noted recently by myself (Tugby, 2024, p. 38), but I did not explore the details. This is something that I aim to remedy here. Importantly, the parallels between the two debates are not merely of historical interest. For example, those working on the topic of goal-directedness might be able to learn something from how the dispositions debate developed after reductive analyses started to fall out of favour in the 1990s. For one thing, there was a resurgence in the Aristotelian idea that dispositions are irreducible properties or 'powers' of things. In the case of goal-directedness, a parallel move would be to accept that goal-directedness is a genuine teleological property had by certain complex systems. Due to the problem of failure, this is an idea that should at least be explored. Moreover, given the parallels between attempts to reduce dispositions and goal-directedness in counterfactual terms, one might wonder whether the two anti-reductive moves—regarding irreducible powers and irreducible teleology—are closely related in some way. For example, perhaps goal-directedness just is a (kind of) power. We shall briefly explore some possibilities in section 5. While it is not my aim to settle all of the questions raised, I hope to show that those working on teleology, and theories of goal-directedness specifically, may benefit from fruitful dialogue with those working on dispositions (and vice versa).

Let us begin by looking at the system-property theory in more detail.

2 | THE SYSTEM-PROPERTY THEORY OF GOAL-DIRECTEDNESS

As noted already, the system-property theory aims to provide an analysis of goal-directedness that is, broadly speaking, naturalistic, behaviouristic, and reductionist. That is to say, it aims to ground truths about goal-directedness in patterns of behaviour that can be adequately described in mechanistic, non-teleological terms. Although the theory acknowledges that talk of goal-directedness might well be useful or even epistemically indispensable in many branches of the special sciences, such talk is not fundamental: in principle, the true scientific theories could be expressed in entirely mechanistic, non-teleological terms. The same point applies to related teleological terms such as 'function'. Many philosophers in the system-property tradition, such as Nagel (1979), think that the concept of goals is important, in part, because *functions* depend on goals (see also Boorse, 1976). Roughly, the idea is that if an item performs a function, it does so because it contributes to some *goal* of a system. For example, one might say that the function of human shivering in cold conditions is to contribute to the goal of maintaining a core body temperature of 37°C. I shall not discuss this goal-theoretic theory of function here (for details see Tugby, 2024, pp. 15–22). The point is just that if this is the correct theory of function, and if talk of goals can be reduced in the way that the system-property theory suggests, then functions themselves will be reducible in non-teleological terms: to say that some item contributes to a goal of a system is just to say that the item contributes to some systemic activity that exhibits a certain sort of complexity. Shortly, we shall look at how system-property theorists have tried to characterise this complexity.

Another salient feature of the system-property theory is that it does not aim to provide an extrinsic, historical theory of goal-directedness. We saw earlier that many goals or functions in biology are grounded extrinsically by causal-historical facts about what the item in question has been selected for. However, for the system-property theorists, natural selection is not an essential feature of goal-directedness. First and foremost, goal-directedness is regarded as a 'here-and-now' property of a system that is grounded in its intrinsic mechanistic organisation. Of course, there will no doubt be important causal-historical stories to be told about why many of these systems are intrinsically organised in the way that they are. In biology, natural selection will play a prominent role in this regard, while in the case of cybernetic systems, the historical story will typically have a lot to do with the intentions of their designers (rather than natural selection). But for system-property theorists, such intentions are neither necessary nor sufficient for conferring the specific goal-directedness that a cybernetic system has. For example, even if a system came about by accident rather than design, it could

arguably still be goal-directed providing that it is 'directively organised' (more on this below). Moreover, the intentions of a designer are not sufficient for goal-directedness in such cases because, arguably, there could be cases in which a system exhibits goal-directed behaviour that is different to what was intended. For instance, Rosenblueth & Wiener point out that a weapon that someone designs could, contrary to their intentions, be directively organised to kill innocent victims (1950, p. 318).

On reflection, it is not at all surprising that system-property theorists tend to deny that there is an essential link between the goal-directedness of a cybernetic system and the intentions of its designers. Again, the aim of the system-property theory is to provide a non-teleological mechanistic reduction of the teleological notion of goal-directedness. But as I have noted elsewhere (Tugby, 2024, p. 27) if we had to ground the goal-directedness of a cybernetic system in the intentions of its designer, we would be merely explaining one teleological notion in terms of another (namely, that of *intention*).

Let us now outline some of the main features of the system-property analysis.[2] We have said that system-property theorists ground goal-directedness in the inherent 'directive organisation' of a system, but what does directive organisation amount to? Different versions of the system-property theory differ in detail, but the basic idea behind directive organisation can be summarised as follows:

> **Directive Organization**: A system is directively organized *if and only if* it displays behaviour that is flexible ('plastic') and highly adaptable ('persistent') with respect to a specific type of outcome.

As E. S. Russell puts it, the hallmark of goal-directed behaviour is 'the active persistence of directive activity towards its goal, the use of alternative means towards the same end, the achievement of results in the face of difficulties' (1945, p. 144). Some system-property theorists then add details about the sorts of mechanisms that give rise to such behaviour. For example, in the early work of Rosenblueth et al. (1943), teleological behaviour is closely associated with negative feedback mechanisms. Such mechanisms continually modify their behaviours in response to environmental input, creating a causal loop of output and input. In the case of *negative* feedback, the input can restrict the output, as in the case of a thermostat, which has a dampening effect if a certain temperature threshold is exceeded.

It is noticeable that E. S. Russell uses teleological terms like 'goal' and 'end' when characterising goal-directed behaviour, but according to system-property theorists, such terms are ultimately dispensable. The idea is that the variability and adaptability involved in goal-directed behaviour can be spelt out in a non-teleological way using the core notions of *plasticity* and *persistence* (e.g., Nagel, 1979, p. 286). These notions can be summarised as follows:

> **Plasticity:** A system is *plastic* with respect to a specific outcome *if and only if* it achieves that outcome from different starting positions or via alternative causal paths.

Persistence, on the other hand, relates to the ongoing adaptability of the behaviour:

> **Persistence:** A system is *persistent* with respect to a specific outcome *if and only if* it is such that if there are any external or internal *disturbances* within a certain range, the system makes *compensatory changes* so that the system maintains its causal path towards the relevant outcome.

A 'disturbance' is that which, in the absence of compensatory changes made by the system, would prevent the realisation of the relevant outcome.[3] To return to our falcon example, it is clear that a falcon's prey-seeking behaviour

[2]For a more detailed and rigorous analysis of the system-property approaches, see Tugby (2024, pp. 28–34).
[3]For some formal characterizations of plasticity and persistence, see e.g., Braithwaite (1953, pp. 329–331) and Nagel (1961, pp. 411–18).

is typically both plastic and persistent. Regardless of where a falcon is placed in, say, a grassy field, it will make its way towards any prey in its vicinity. The falcon's behaviour is therefore plastic with respect to catching prey. Moreover, if its prey moves, or if the falcon encounters an obstacle enroute, it will adjust its flight accordingly. The falcon can therefore compensate for disturbances and exhibit persistence. According to the system-property theory, then, the falcon's behavioural profile has all the hallmarks of a directively organised system vis-à-vis prey catching. And mutatis mutandis for other goal-directed systems, including those which maintain some equilibrium state rather than seek out a target. For example, the human heat moderation system is goal-directed because it works well in different environments (plasticity) and can compensate for a good range of disturbances (persistence), such as a sudden drop or rise in external temperature, to maintain the subject's body temperature of 37°C.

In summary, in its most basic form the system-property theory states that a system is directively organised or 'goal-directed' if and only if it exhibits plastic and persistent behaviour with respect to a certain outcome. And notice that, above, we were able to characterise plasticity and persistence using entirely causal, non-teleological language. Hence, the system-property theory appears to achieve its reductive aims. It is also thoroughly empirical given that patterns of plastic and persistent behaviour, and the mechanisms which give rise to such behaviour, can typically be observed and tested.

3 | THE PROBLEM OF FAILURE AND A COUNTERFACTUAL RESPONSE

There is, however, a well-known problem facing various versions of the system-property analysis, which has been highlighted by numerous authors (e.g., Ehring, 1984a; Nissen, 1997; Scheffler, 1963; Tugby, 2024, pp. 35–36). As Scheffler puts it, due to the behaviourist roots of the system-property analysis, goal-directedness can only be ascribed to a given type of system 'on the basis of observation of the conditions under which similar behavior has taken place in the past' (1963, p. 118). More precisely, goal-directedness may be ascribed to a certain kind of system once we have observed trials and witnessed such systems achieve an outcome via various causal paths and in the face of various disturbances. This means that the system-property theories of Braithwaite and others are what Nissen calls 'trial and success' theories (1997, p. 7). The problem, though, is that it is unclear whether and how we can ascribe goal-directed behaviour in cases where the putative goal is rarely, if ever, reached. As Nissen puts it, 'No provision is made for failing to realize the goal-state … It ignores the large amount of behavior directed to goals never reached' (1997, p. 7). Yet surely there can be goal-directed behaviour towards goals which are rarely or never reached. For example, in so-called cases of 'undamped feedback', a goal-directed machine might consistently overshoot its target and never recover (Rosenblueth et al., 1943, p. 20). We might call these the 'failure cases'. How can the system-property theory handle such cases? What is the difference between a system that is goal-directed but never achieves its goal, and one that is not goal-directed at all?

When raising this problem, Scheffler (1963, p. 119) notes that Braithwaite could begin to address this problem by dropping the requirement that the relevant causal chains actually achieve the relevant goal. However, this quickly leads to further problems. First and foremost, it is difficult to see what resources the system-property theorists can use to determine the goal of a system in those cases where the goal fails to be achieved. Scheffler calls this the 'difficulty of multiple goals' (1963, p. 120), because in an indefinitely large number of failure cases, there will be many different competing hypotheses regarding the goal of the behaviour—hypotheses which are all consistent with what is observed. We could equally call this a problem of underdetermination; the worry being that in the failure cases, the overt behaviour of a system underdetermines the goal of that behaviour (Tugby, 2024, p. 35). To use Scheffler's example, suppose that a cat is crouching in front of a mousehole, but nothing emerges from the hole (1963, p. 120). On what basis can we say that the cat's behaviour is directed towards catching a mouse rather than, say, a bird? The crouching behaviour is consistent with either of these goal hypotheses. At this point, one might of course be tempted to say in such cases that the goal is grounded in, and made determinate by, the specific *intentions* of the organism. Alternatively, in the

case of non-biological goal-directed artefacts, one might suppose that the goals of the system are grounded by the intentions of the system's designer(s). But as we saw in the previous section, these are not options that system-property theorists will want to take, not least because they would end up appealing to further teleological concepts (such as that of *intention*)—concepts that they ultimately want to dispense with.

There appears to be a fairly obvious fix that system-property theorists could try to employ. Commentators such as Woodfield (1976, p. 49) observed that the most obvious lesson for system-property theorists to learn is that they need to analyse goal-directedness in modal *counterfactual* terms, rather than in terms of actual patterns of behaviour. Although Scheffler's cat might not *actually* succeed in catching a mouse, we may nonetheless point out that *if* a mouse had appeared in some way, then the cat *would* have pursued it (in a persistent way, as required). The cat's goal would then be made determinate not by actual patterns of behaviour but rather *counterfactual truths* about such behaviour. Importantly, this account would still be reductive if the relevant counterfactuals are formulated in a way that does not employ any teleological terms such as 'goal' or 'intention'.

I concur that a counterfactual formulation of goal-directedness probably offers the best chance of success for reductive system-property theorists.[4] And indeed, some formulations of the system-property theory have been couched in counterfactual terms. For example, Sommerhoff (1950, Ch. 2) defines 'adaptation' and 'directive correlation' in counterfactual terms. And when discussing specific examples of goal-directedness in biology, such as the human water moderation mechanism, Nagel finds it natural to employ counterfactual language (1979, p. 287).[5]

Unfortunately, though, there remains a serious problem, which is that the most obvious counterfactual formulations of goal-directedness face counterexamples (Tugby, 2024, pp. 27–38). In the next section, we shall see that this debate regarding counterfactual reduction plays out in much the same way as the dispositions debate, where reductionists have attempted to provide counterfactual analyses of dispositions. The various formulations, proposed in the 1990s onwards, face a variety of counterexamples which closely resemble those that are scattered throughout (earlier) critical discussions of the system-property theory. It is far from clear that a non-circular, non-ad hoc, and counterexample-free counterfactual analysis of dispositions has been found. And given the close but overlooked parallels between reductive counterfactual approaches to dispositions and goal-directedness, the prospects for a problem-free counterfactual analysis of goal-directedness do not look especially good. After drawing out the parallels between the two debates in the next section, we shall consider what general lessons might be learnt by those seeking a general philosophical theory of goal-directedness.

4 | COMPARING COUNTERFACTUAL ACCOUNTS OF GOAL-DIRECTEDNESS AND DISPOSITIONS

4.1 | Simple analyses and the problem of finks

Dispositions are behavioural properties that are essentially tied to certain sorts of causal outcome: properties such as elasticity (the disposition to stretch), fragility (the disposition to break), or solubility (the disposition to dissolve). A 'simple' counterfactual analysis of dispositions is one that equates *x*'s disposition to manifest M (when stimulated by S) with the following counterfactual:

[4]A rather different option is to insist that goal-directed systems always possess internal representations of their goal states, and that the goal of a system is determined by those representations even if the goal is never achieved. This idea is explored in Adams (1979) and Faber (1986). The obvious problem with this proposal is that it is difficult to see how simple goal-directed systems can literally exhibit inner representational states. For critical discussion of this representation strategy, see e.g., Ehring (1984b), Nissen (1997), Tugby (2024, pp. 39–40) and Woodfield (1976).
[5]More recently, Stovall (2024) has provided a detailed discussion of the subjunctive aspects of biological plasticity and persistence.

> **Simple Analysis of Dispositions:** *x* has the disposition to M when S'ed *if and only if*, if *x* were to undergo stimulus S, then *x* would M.

Due to the work of C. B. Martin, it is well known that the simple counterfactual analysis of dispositions faces counterexamples involving what Martin calls 'finks' (1994, p. 2). Let us illustrate with Martin's own example of a wire's dispositional property of being live. One might think that this dispositional property can be analysed straightforwardly in the following way:

> **Live:** *x* is live *if and only if*, if *x* were in contact with a conductor, then electric current would flow through the conductor (adapted from Martin, 1994, p. 2).

In response, Martin offers a counterexample involving an 'electro-fink' machine (1994, p. 3), which switches off the power supply at the instant at which the wire is touched by the conductor. In such a case, the counterfactual on the right-hand side of the analysis above is rendered false because it would not be the case that electric current flows when the wire is put in contact with a conductor. And yet, we should surely not conclude from this that the wire in question was not live. *Ex hypothesi*, the wire was indeed live before it was touched, which means that the simple counterfactual above has delivered the wrong verdict. Hence, the disposition ascription and the simple counterfactual cannot be equivalent.

Interestingly, the same sorts of arguments unfold when we consider a simple counterfactual analysis of goal-directedness, such as the cat and mouse counterfactual introduced in the previous section. *Ex hypothesi*, the crouching behaviour of Scheffler's cat is directed towards the goal of catching a mouse, even if a mouse never appears. We might try to unpack this using the following simple counterfactual:

> **Cat Goal:** The cat's crouching behaviour is directed towards the goal of catching a mouse *if and only if*, if a mouse were to appear from the hole, then the cat would have caught it (via persistent means, as required).

The 'via persistent means' clause is there to capture the system-property theorists' plausible insight that, as a goal-directed system, the cat would adapt its behaviour accordingly to catch the mouse. For example, if, in the counterfactual scenario, the mouse tries to run away, the cat would chase after it before pouncing. But as far as the counterfactual is concerned, the precise details regarding the mouse-capturing mechanism are irrelevant. What is important, as far as the problem of failure is concerned, is that the counterfactual specifies the outcome of the process and thereby specifies the goal of the cat's behaviour.

Unfortunately, though, it is not difficult to see that fink-type counterexamples can also be concocted for this simple analysis of goal-directedness. For instance, suppose that at the moment at which the mouse appears, the cat's owner would pass some healthier food that the cat much prefers (some succulent salmon). Thus, in the counterfactual scenario, the cat loses the desire to catch the mouse and focuses their attention on the salmon instead. In that case, the appearance of a mouse does not lead to the cat catching the mouse. And yet, *ex hypothesi*, the cat's crouching behaviour was indeed goal-directed towards the catching of a mouse. Hence, like Martin's electrofink case, this example shows that a simple counterfactual analysis will not always deliver the correct verdict.

In earlier critical discussions of the system-property theory, similar counterexamples can be found which predate Martin's finkish cases by at least a decade. For example, when discussing Nagel's theory of goal-directedness, Nissen (1981, p. 133) considers a hypothetical case in which someone genuinely has the goal of unlocking a door but would 'give up' if the lock had an intimidating appearance. Nissen also mentions another example involving someone who has the goal of lifting a heavy weight but is susceptible to injury: they would quickly pull a muscle and be unable to carry out the lifting of the weight (Nissen, 1981, p. 133). Such cases are

presented as a problem for Nagel because, *ex hypothesi*, the subjects have the goals in question, but counterfactually they would not display the relevant persistent behaviour and would not achieve the relevant goals. It is not difficult to see that these counterfactual scenarios have a similar structure to those that Martin concocts in the dispositions debate. In these cases, we might say that the goal-directed behaviour is 'finked'. The worry facing system-property theorists is that no matter how plastic and persistent a goal-directed system is, it seems possible to cook up finkish obstacles that cause the goal-directedness to be lost as soon as the goal-directed process is initiated.

Unsurprisingly, reductionists about dispositions tried to modify their counterfactual analyses by inserting a proviso, one whose role it is to ensure that the relevant disposition is not lost before the manifestation occurs. Without going into unnecessary details, Lewis (1997) tries to do this by pointing out that finks generally operate by altering an intrinsic property of the subject—namely, the property that is the *causal base* of the disposition. Lewis then uses this insight to modify the counterfactual analysis. In order to exclude the presence of finkish interferences, the antecedent of Lewis's counterfactual schema specifies that the relevant intrinsic property (the causal base of the disposition) is *retained* for a certain amount of time—namely, the amount of time it takes for the relevant disposition to manifest once stimulated.[6] Possible worlds where finkish interferences occur then become irrelevant for the assessment of the counterfactual, because a finkish world is one in which the causal base of the disposition is *not* retained.

On the surface, Lewis's solution looks promising, and it is the kind of solution that, prima facie, could be implemented by system-property theorists in the case of goal-directedness. Given the mechanistic nature of their account, the system-property theorists would surely accept that goal-directed behaviour has a mechanistic causal base. They would also agree, at least typically, that such bases are intrinsic (see e.g., Rosenblueth et al., 1943, p. 19). One could therefore try to rescue the counterfactual analysis of goal-directedness by inserting a Lewisian proviso stating that the relevant intrinsic properties of the goal-directed subject are retained for a certain amount of time. In the cat and mouse case, the relevant intrinsic properties might (for example) include certain psychological features. The modified counterfactual analysis might then take something like the following form:

> **Cat Goal (Reformed):** The cat's crouching behaviour at *t* is directed towards the catching of a mouse *if and only if*, if a mouse were to appear from the hole at *t*, and the cat were to retain certain psychological and physical properties until *t′*, then the cat would have caught the mouse (via persistent means, as required).

One may, of course, have further questions about the details of formulation in this case. For example, one might wonder how the relevant psychological or physical properties are to be specified, and how long the period between *t* and *t′* would have to be. But those details need not concern us, because however those details play out, there remains a problem. The problem is that other kinds of counterexample will apply even if the Lewisian modifications are adopted. In the dispositions literature, these include counterexamples involving so-called 'antidotes' and 'mimickers'. Such counterexamples were raised by numerous critics in response to modifications made by Lewis and other reductionists about dispositions. As we shall see, analogous problem cases can also be found scattered throughout earlier critical discussions of the system-property theory of goal-directedness.

[6]For completeness, here is Lewis's reformed analysis of dispositions:

> Something *x* is disposed at time *t* to give response *r* to stimulus *s* iff, for some intrinsic property *B* that *x* has at *t*, for some time *t′* after *t*, if *x* were to undergo stimulus *s* at time *t* and retain property *B* until *t′*, *s* and *x*'s having of *B* would jointly be an *x*-complete cause of *x*'s giving response *r* (1997, p. 157).

4.2 | Antidotes and mimickers

Antidote cases are those in which the subject *does* retain the intrinsic causal base of the disposition in the counterfactual scenario, but where some extrinsic interfering factor prevents the disposition from achieving the expected manifestation. One of the first examples of a dispositional antidote, proposed by Bird (1998, p. 229) in response to Lewis (1997), concerns a nuclear uranium pile which has the disposition to produce a devastating chain reaction upon reaching a certain critical mass. Despite this disposition, it turns out that even if a uranium pile goes above critical mass, one can prevent the chain reaction by inserting boron rods which absorb the radiation. That is a fortunate fact for humans who rely on nuclear energy, but less fortunate for Lewis. Importantly, the boron rods appear to perform this preventative role *without* altering the intrinsic properties of the uranium pile. The boron rods are therefore not finks in Martin's sense and hence are not excluded by the Lewisian modification discussed earlier. Hence, Lewis-style dispositional counterfactuals do not appear to deliver the correct verdict in cases of antidotes. *Ex hypothesi*, it is true that uranium piles have the disposition to chain-react, but the corresponding Lewisian counterfactual is rendered false in cases where boron safety mechanisms are in place.

Cases involving so-called mimicry also provide counterexamples to various counterfactual analyses of dispositions, but for the opposite reason. While an antidote prevents the manifestation of a disposition that a subject really does have, a mimicker produces an effect for which the subject does not have a corresponding disposition. So, while antidote cases purport to show that counterfactual analyses do not accommodate enough dispositions, cases of mimicry suggest that counterfactual analyses over-generate dispositions in some cases. One of the first examples of mimicry came from Lewis himself (1997, p. 153). In the example, there is a styrofoam dish which, *ex hypothesi*, is not fragile. However, if the dish were struck, it would make a distinctive noise that annoys someone in the vicinity ('Hater of Styrofoam'), who proceeds to rip the dish apart. So, it is true to say of the dish that if it were struck, then it would break, and yet *ex hypothesi* the dish is not fragile.

Once again, alleged counterexamples can be found in earlier discussions of goal-directedness that precisely mirror the problem cases of antidotes and mimickers. For example, in a critique of Nagel's system-property theory, Ehring briefly considers an example involving a bird of prey which 'systematically misses its target, generally only wounding rather than killing' (1984a, p. 502). Given that this behaviour is systematic, the counterfactual (as well as actual) behaviour of the bird misleading suggests that its goal is to wound prey, when *ex hypothesi* the real goal is to kill prey. Interestingly, this example could be fleshed out either as a case of mimicry or in terms of antidotes. It could be viewed as a case of goal mimicry because whenever the bird encounters prey, it acts persistently in a way that leads to the non-fatal wounding of the prey (rather than killing). So, this bird of prey always behaves *as if* the goal is to non-fatally wound, when in fact that is not really the goal that it has. However, we could also develop the case in a way that involves an antidote. For example, suppose that the bird's environment is one in which the prey regularly casts unusual shadows which disorientate the bird. In that case, we might think of the shadows as antidotes or masks for the bird of prey's goal-directedness.

When Scheffler discusses his cat counterexample, he develops it in a way that could also be regarded as a case of goal mimicry. The cat is crouching in front of the mousehole with the specific goal of catching a mouse (rather than, say, catching some other kind of food). However, Scheffler's worry, recall, is that system-property theorists do not always have the resources to determine that this is the cat's goal rather than some other. It is, let us suppose, true that if a mouse were to appear, then the cat would pursue it persistently. But as Scheffler points out, it is no less true that the cat would pursue a bowl of cream, if such a bowl were to appear. As Scheffler puts it, there are 'hypothetical sets of field conditions, each set including one positing a bowl of cream within the cat's range, such that, conjoined to the cat's present behaviour, each set determines a cream-attaining causal chain' (Scheffler, 1963, p. 120). The essential point here is that in this hypothetical

scenario, the cat behaves as if its goal is to get a bowl of cream, even though this was not in fact the goal of its crouching behaviour. Hence, the worry is that counterfactual analyses of goal-directedness will over-generate goals in some cases.

Now, those who examined the counterfactual approach to goal-directedness were not unaware of the various problem cases. For example, when considering a counterfactual formulation of Braithwaite's plasticity theory of goal-directed behaviour, Woodfield proposes the following provisos:

> If obstacles O_x had not been present in field-conditions f_x, which otherwise fall within the variancy, or if an essential member of the set f_x had not been absent (we may call this an obstacle too), the behaviour would have ended in Γ
>
> (1976, p. 49).

Prima facie, the provisos that eliminate obstacles are just what are needed to avoid many of the counterexamples discussed above, which involve interfering factors of one sort or another. Woodfield was clearly aware that any successful counterfactual analysis of goal-directedness must include provisos ruling out factors that would frustrate the attainment of the goal.[7] However, as the subsequent dispositions debate has taught us, it is far from easy to flesh out such provisos in a way that is non-circular and non-trivial (e.g., Martin, 1994, p. 6; Tugby, 2024, p. 38). For example, for Woodfield's proposal to be informative and precise, we need to be able to grasp what counts as an 'obstacle'. But clearly, if we had to spell out the proviso by listing *every possible* circumstance that would frustrate the goal-attaining process, of which there are potentially indefinitely many, then the analysis would become impossible to formulate. On the other hand, we cannot simply define an obstacle as that which would prevent the outcome from occurring, for then Woodfield's provisos would be trivial. They would, in effect, be saying something like 'the outcome Γ would occur, unless something happens which makes Γ not occur'. Provisos of this form are trivially true; in which case they surely cannot be used as part of an illuminating analysis of goal-directedness.

Ever since these problems came to light in the dispositions debate, there has been a philosophical cottage industry of modifications to the counterfactual analysis of dispositions, followed by further counterexamples.[8] Although the debate is ongoing, many of us suspect that there is already enough evidence to suggest that a reductive counterfactual analysis is the wrong approach for theorising about dispositions. And given the parallels that we have drawn between the reductive dispositions and goal-directedness debates, one might well come to a similar conclusion about the counterfactual analysis of goal-directedness.

Nonetheless, it is not my aim here to argue conclusively that a satisfactory counterfactual analysis of goal-directedness cannot be formulated. For current purposes, the main point is just that due to the difficulties that the reductive approach faces, it is surely a worthwhile exercise to consider what other approaches we might take when theorising about goal-directedness. This is the business of the rest of the paper. Again, we shall see that there might well be lessons to be learnt from the dispositions debate. In the light of the difficulties facing the reductive analyses of dispositions, many dispositions theorists took a realist turn and began taking seriously the idea that many dispositional properties—or what are often called 'powers'—are real, irreducible features of the world. According to the realist powers view, although certain counterfactual truths might be *symptomatic* of certain dispositions, those counterfactuals do not *constitute* the dispositions; insofar as there is a relation of determination between dispositions and certain counterfactuals, it is the former that determine the latter rather than vice versa. In the case of goal-directedness, the parallel idea would be that certain biological and non-biological systems

[7]To be clear, Woodfield is not himself a system-property theorist. In this passage, he is merely considering how someone like Braithwaite could try to deal with the obvious problem cases.

[8]One of the more promising recent proposals regarding the link between dispositions and conditionals has been offered by Manley and Wasserman (2008, 2011). Their account employs the notion of proportions: '*N* is disposed to *M* when *C* iff *N* would *M* in a suitable proportion of *C*-cases' (2008, p. 76). For discussion of some problems facing this account, including another triviality worry, see Vetter (2011).

have genuine properties of goal-directedness which determine certain counterfactual truths about those systems (rather than vice versa).

5 | A REALIST TURN

In response to the perceived difficulties facing reductionism about dispositions, Martin (1994), Molnar (2003), Mumford (2004) and others have offered a rather different approach, which involves accepting the Aristotelian idea that dispositions or powers are real, irreducible features of things. There are different ways of thinking about powers (more on this below), but the variants share the idea that powers are essentially directed towards, and at least partly individuated by, the manifestation types that they are powers *for*. Powers theorists also accept that powers are in some sense directed towards their manifestations regardless of whether those manifestations occur. For these reasons, powers are often regarded as being teleological in some sense (see e.g., Kroll, 2017; Oderberg, 2017, 2020; Tugby, 2020; Witt, 2008).

Given the similarities between the dispositions debate and the corresponding debate about goal-directedness, one would not be surprised to see a parallel resurgence of the idea that the goal-directedness of certain entities is a genuine teleological, end-directed state. Indeed, this resurgence has already started to take shape, with several notable works proposing thoroughgoing realism about teleological properties (e.g., Austin & Marmodoro, 2017; Feser, 2014; Kroll, 2017; Oderberg, 2017, 2020; Page, 2021; Paolini Paoletti, 2021; Tugby, 2024; Witt, 2008, to name a few). In the current context, a realist stance would involve the idea that various counterfactual truths about directively organised systems hold *because* of the systems' goal-directedness rather than vice versa.

Those who are realists about powers might go a step further and take the relevant parallels as an indication that real states of goal-directedness just are powers of a certain sort (e.g., Tugby, 2024, sect. 4). Indeed, even Nagel's system-property theory of goal-directedness has sometimes been interpreted as a dispositional view—albeit in terms of dispositions that are supposed to be analysable in counterfactual terms (e.g., Boorse, 2002, p. 69; Faber, 1986, p. 61). But if one adopts a strong realism about dispositions, one might wonder whether states of goal-directedness just are powers of a certain sort. We could even develop such a view by drawing on some aspects of the system-property theory. We might, for instance, characterise states of goal-directedness as powers of complex systems that have a *plastic* and *persistent* modal profile (Tugby, 2024, pp. 48–49). Such powers would be plastic in the sense that their instances could be activated at different starting positions and in different environments; and they would be persistent in the sense that their possessors can find a way to manifest the power when faced with certain obstacles and environmental disturbances.[9] There would, of course, be further detailed questions to ask about how plastic and persistent a power has to be in order for it to count as goal-directed.[10] If the threshold is low, then most or even all powers might be regarded as goal-directed.[11]

Importantly, cases of goal failure would not present an obvious problem for a powers-based theory of goal-directedness.[12] As noted above, powers are, by their very nature, directed towards a determinate manifestation-type regardless of whether the powers are actually manifested. There may be many powers for manifestations

[9]Such powers might, of course, be grounded by further lower-level powers of the system. However, grounding is not the same thing as reduction: see Tugby (2022, ch. 6.2) for discussion.

[10]It is well worth noting that because dispositions appear to be gradable (Manley & Wasserman, 2007), a dispositional account of goal-directedness should be well placed to accommodate the plausible idea that goal-directedness comes in degrees (see e.g., Tugby, 2024, p. 60).

[11]Cartwright (2019, ch. 2), for one, seems happy to regard all powers as being goal-directed in some sense. Recently Babcock (2023) also argues that goal-directedness occurs throughout the non-living realm, although following McShea (2012) he expresses the idea in terms of 'forces' rather than powers.

[12]That is not to say that the powers-based theory does not face any problems. For example, as I have noted elsewhere (Tugby, 2024, sect. 4.8), it is not entirely clear whether a powers-based account of goal-directedness can accommodate apparent cases of goal-directedness in which the goal is impossible to achieve. Unfortunately, I do not have the space to explore all the various ramifications of the powers-based view. I welcome further work on these issues.

that are never produced. Examples of goal failure would just be special cases of this phenomenon. Even if the relevant goal is not achieved, the goal-directedness of a system can nevertheless be grounded by the relevant plastic and persistent power(s).

There are different ways in which the essential directedness of powers could be understood, and therefore different ways of spelling out the metaphysical nature of goal-directed powers. I will mention just three of the prominent variants here. Firstly, one of the early accounts in the contemporary powers movement is based on the idea that powers have all the hallmarks of *intentionality* (e.g., Bauer, 2022; Martin & Pfeifer, 1986; Molnar, 2003; Place, 1996). In the same way that intentional thoughts can be directed towards intentional objects that do not exist, so too powers can be directed towards manifestations that are never produced. Intentionality theorists take this analogy metaphysically seriously and regard physical powers as states of physical intentionality. In the case of goal-directed powers, the relevant goal would be determined by a state of physical intentionality that the directively organised system instantiates.

Other powers theorists have offered an account of powers that is *relational* rather than intentional. Relational powers theorists argue that to make metaphysical sense of a power's directedness, we need to regard such directedness as consisting in a genuine asymmetric relation between properties. Bird (2007, ch. 6), for instance, calls such relations 'stimulus-manifestation' relations, while I have previously called them relations of 'dispositional directedness' (Tugby, 2022, ch. 2). Given that such relations hold between the properties themselves rather than individuals, they are *second-order* relations. And given that a power's directedness is essential, such relations help to individuate and constitute the powers. Powers would thus compose an essential (second-order) relational network that Bird represents using the resources of graph theory (2007, ch. 6.; see also Tugby, 2013). In the case of goal-directed powers, then, a relational theory would represent such powers as nodes in a graph which are related to their goal-types (and possibly their stimulus-types) via the relevant second-order relations.

Thirdly, Kroll (2017) and others have recently proposed an overtly teleological theory of powers, arguing that it is superior to previous accounts of dispositions in various respects. According to Kroll's teleological analysis, powers are identified with primitive teleological states of end-directedness. As Kroll puts it, 'Necessarily, the property of being disposed to *M* when *C just is* the property being in a state directed at the end that one *Ms* when *C*' (2017, p. 21). Within this framework, a goal-directed power would simply be an end-directed telic state that is instantiated by a system of the right sort of complexity. The goal would be whatever end it is that the system's telic state is directed towards.[13]

My aim in this section has been to map out some of the theoretical terrain rather than to determine which of these specific accounts of goal-directed powers is the best. A more detailed critical assessment is provided in Tugby, 2024, sect. 4. However, as regards the problem of goal failure, what seems clear is that all the proposals above provide resources for determining the relevant goal(s) in cases where the goal is not achieved. Powers are tailormade for this role because they are directed entities by their very nature. And given that powers occur in all realms of nature, the theory has broad applicability, covering the range of both biological and non-biological cases.

6 | CONCLUSIONS

We have examined the reductive system-property theory of goal-directedness and found reasons for thinking that it needs to be formulated in counterfactual terms. However, it proves to be difficult to formulate a counterfactual analysis of goal-directedness that is counterexample-free, non-circular, and non-trivial. These difficulties closely mirror those facing reductionists about dispositions, though the parallels between the two

[13]See Tugby (2024, sect. 4.3), for a more detailed discussion of how a Kroll-type theory of dispositions can be applied to cases of goal-directedness.

debates have been overlooked in the literature. After outlining some of the parallels, we considered what goal theorists might learn from the dispositions debate. We explored the need for a realist, non-reductionist account of goal-directedness, and considered the idea that properties of goal-directedness are themselves dispositions or 'powers' of a certain sort. Overall, I hope to have shown that those working on teleology, and theories of goal-directedness specifically, may benefit from fruitful dialogue with those working on dispositions and powers (and vice versa).

ACKNOWLEDGEMENTS

Versions of this paper were presented at the Dispositions and Powers Conference at the University of Bristol (organised by Toby friend and Samuel Kimpton-Nye), the Purpose in Biology: New Directions Conference at the University of Reading (organised by David Oderberg and Christopher Austin), and the Penelope Mackie Memorial Conference at the University of Nottingham (organised by Neil Sinclair). I am very grateful to the organisers of those events and to the audience members for their valuable feedback. This work is dedicated to my late father, who passed away as I was completing it.

ORCID

Matthew Tugby https://orcid.org/0000-0001-5199-5982

REFERENCES

Adams, F. R. (1979). A goal-state theory of function attributions. *Canadian Journal of Philosophy*, *9*, 492–518. https://doi.org/10.1080/00455091.1979.10716265

Austin, C. J., & Marmodoro, A. (2017). Structural powers and the homeodynamic unity of organisms. In W. M. R. Simpson, R. C. Koons, & N. J. Teh (Eds.), *Neo-Aristotelian perspectives on contemporary science* (pp. 169–184). Routledge. https://doi.org/10.4324/9781315211626-10

Babcock, G. (2023). Teleology and function in non-living nature. *Synthese*, *201*(112), 1–20. https://doi.org/10.1007/s11229-023-04099-1

Bauer, W. A. (2022). *Causal powers and the intentionality continuum*. Cambridge University Press.

Bird, A. (1998). Dispositions and antidotes. *Philosophical Quarterly*, *48*, 227–234. https://doi.org/10.1111/1467-9213.00181

Bird, A. (2007). *Nature's metaphysics: Laws and properties*. Oxford University Press.

Boorse, C. (1976). Wright on functions. *Philosophical Review*, *85*, 70–86. https://doi.org/10.2307/2184255

Boorse, C. (2002). A rebuttal on functions. In A. Ariew, R. Cummins, & M. Perlman (Eds.), *Functions: New essays in the philosophy of psychology and biology* (pp. 63–112). Oxford University Press. https://doi.org/10.1093/oso/9780199255801.003.0004

Braithwaite, R. B. (1953). *Scientific explanation*. Cambridge University Press.

Cartwright, N. (2019). *Nature, the artful modeler: Lectures on laws, science, how nature arranges the world and how we can arrange it better*. Open Court Publishing Company.

Ehring, D. (1984a). The system-property theory of goal-directed processes. *Philosophy of the Social Sciences*, *14*, 497–504. https://doi.org/10.1177/004839318401400405

Ehring, D. (1984b). Negative feedback and goals. *Nature and System*, *6*, 217–220.

Faber, R. J. (1986). *Clockwork garden: On the mechanistic reduction of living things*. University of Massachusetts Press.

Feser, E. (2014). *Scholastic metaphysics*. Editiones Scholasticae.

Kroll, N. (2017). Teleological dispositions. In K. Bennett & D. W. Zimmerman (Eds.), *Oxford studies in metaphysics* (Vol. *10*, pp. 1–37). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198791973.003.0001

Lewis, D. (1997). Finkish dispositions. *Philosophical Quarterly*, *47*, 143–158. https://doi.org/10.1111/1467-9213.00052

Manley, D., & Wasserman, R. (2007). A gradable approach to dispositions. *Philosophical Quarterly*, *57*, 68–75. https://doi.org/10.1111/j.1467-9213.2007.469.x

Manley, D., & Wasserman, R. (2008). On linking dispositions with conditionals. *Mind*, *117*, 59–84. https://doi.org/10.1093/mind/fzn003

Manley, D., & Wasserman, R. (2011). Dispositions, conditionals, and counterexamples. *Mind*, *120*, 1191–1227. https://doi.org/10.1093/mind/fzr078

Martin, C. B. (1994). Dispositions and conditionals. *Philosophical Quarterly*, *44*, 1–8. https://doi.org/10.2307/2220143

Martin, C. B., & Pfeifer, K. (1986). Intentionality and the non-psychological. *Philosophy and Phenomenological Research*, *46*, 531–554. https://doi.org/10.2307/2107668
McShea, D. W. (2012). Upper-directed systems: A new approach to teleology in biology. *Biology and Philosophy*, *27*, 663–684. https://doi.org/10.1007/s10539-012-9326-2
Molnar, G. (2003). In S. Mumford (Ed.), *Powers: A study in metaphysics*. Oxford University Press.
Mumford, S. (2004). *Laws in nature*. Routledge.
Nagel, E. (1961). *The structure of science*. Harcourt, Brace and World.
Nagel, E. (1979). Teleology revisited. In *Teleology revisited and other essays in the philosophy and history of science* (pp. 275–316). Columbia University Press.
Nissen, L. (1981). Nagel's self-regulation analysis of teleology. *The Philosophical Forum*, *12*, 128–138.
Nissen, L. (1997). *Teleological language in the life sciences*. Rowman and Littlefield.
Oderberg, D. S. (2017). Finality revived: Powers and intentionality. *Synthese*, *194*, 2387–2425. https://doi.org/10.1007/s11229-016-1057-5
Oderberg, D. S. (2020). *The metaphysics of good and evil*. Routledge.
Page, B. (2021). Power-ing up neo-Aristotelian natural goodness. *Philosophical Studies*, *178*, 3755–3775. https://doi.org/10.1007/s11098-021-01624-1
Paolini Paoletti, M. (2021). Teleological powers. *Analytic Philosophy*, *62*, 336–358. https://doi.org/10.1111/phib.12245
Place, U. T. (1996). Intentionality as the mark of the dispositional. *Dialectica*, *50*, 91–120. https://doi.org/10.1111/j.1746-8361.1996.tb00001.x
Rosenblueth, A., & Wiener, N. (1950). Purposeful and non-purposeful behavior. *Philosophy of Science*, *17*, 318–326. https://doi.org/10.1086/287107
Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, *10*, 18–24. https://doi.org/10.1086/286788
Russell, E. S. (1945). *The directiveness of organic activities*. Cambridge University Press.
Scheffler, I. (1963). *The anatomy of inquiry*. Alfred A. Knopf.
Sommerhoff, G. (1950). *Analytical biology*. Oxford University Press.
Stovall, P. (2024). The teleological modal profile and subjunctive background of organic generation and growth. *Synthese*, *203*(77), 1–37. https://doi.org/10.1007/s11229-023-04438-2
Tugby, M. (2013). Graph-theoretic models of dispositional structures. *International Studies in the Philosophy of Science*, *27*, 23–39. https://doi.org/10.1080/02698595.2013.783979
Tugby, M. (2020). Organic powers. In A. S. Meincke (Ed.), *Dispositionalism: Perspectives from metaphysics and philosophy of science* (pp. 213–238). Springer. https://doi.org/10.1007/978-3-030-28722-1_13
Tugby, M. (2022). *Putting properties first: A platonic metaphysics for natural modality*. Oxford University Press.
Tugby, M. (2024). *Teleology*. Cambridge University Press. https://doi.org/10.1017/9781009257404
Vetter, B. (2011). On linking dispositions and which conditionals? *Mind*, *120*, 1173–1189. https://doi.org/10.1093/mind/fzr077
Witt, C. (2008). Aristotelian powers. In R. Groff (Ed.), *Revitalizing causality: Realism about causality in philosophy and social science* (pp. 129–138). Routledge. https://doi.org/10.4324/9780203932636-15
Woodfield, A. (1976). *Teleology*. Cambridge University Press.

How to cite this article: Tugby, M. (2024). The property of goal-directedness: Lessons from the dispositions debate. *Ratio*, *37*, 313–326. https://doi.org/10.1111/rati.12417