Hamilton's rule: A non-causal explanation?

Citation for published version (APA):
Koliofotis, V., & Verreault-Julien, P. (2022). Hamilton's rule: A non-causal explanation? Studies in History and Philosophy of Science, 92, 109-118. https://doi.org/10.1016/j.shpsa.2021.11.008

Document license:
CC BY

DOI:
10.1016/j.shpsa.2021.11.008

Document status and date:
Published: 01/04/2022

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Hamilton’s rule: A non-causal explanation?∗

Vaios Koliofotis a,∗, Philippe Verreault-Julien b,1

a Erasmus University Rotterdam, the Netherlands
b Eindhoven University of Technology, the Netherlands

A B S T R A C T

The explanatory power of Hamilton’s rule, the main explanatory principle of social evolution theory, is an ongoing subject of controversy. In this paper, we reinforce the case for the considerable value of the regression-based version of the rule in explaining the evolution of social traits. Although we agree that the rule can have an organizing role in social evolution research, we maintain that it does not explain in virtue of citing causes or providing an organizing framework. Instead, we argue it either provides an explanation by constraint or a non-causal counterfactual explanation.

1. Introduction

Ever since Darwin (1859, 1871), biologists have been interested in providing evolutionary explanations of social traits. Hamilton’s (1963, 1964a,b) groundbreaking work showed why costly social traits like altruism could evolve by natural selection. Although it was not clear what were the model’s limitations and whether its assumptions were plausible, half a century of research has consolidated Hamilton’s solution. Challenges from population genetics (e.g. Cavalli-Sforza & Feldman, 1978; Karlin & Matsessi, 1983; Matsessi & Karlin, 1984) led to generalizations of inclusive fitness theory (IFT henceforth) (e.g. Queller, 1992), to the examination of its relation to genetics (e.g. Karlin & Matsessi, 1983), and to evolutionary game theory (e.g. Taylor & Frank, 1996).

Despite the immense theoretical progress and the development of sophisticated empirical methods during that same period, social evolution theory continues to be controversial. Notably, high-profile biologists Nowak et al. (2010) joined forces in criticizing IFT, especially as it applies to microbes and social insects. Their Nature article provoked a fierce response of more than one hundred evolutionary biologists (Abbot et al., 2011) and subsequently generated a growing body of highly technical literature (e.g. Allen et al., 2013; Frank & Fox, 2020; Gardner et al., 2011; Lehmann & Rousset, 2014; Nowak et al., 2017; Rousset, 2015; van Veelen et al., 2017).

The exchange has also taken a philosophical twist beyond mathematical modelling or empirical research. Philosophers of science have attempted to clarify conceptual and methodological misunderstandings that arise from different formulations of social evolution theory’s main explanatory principle, Hamilton’s rule (HR henceforth) (Birch, 2014, 2017a,b; Birch & Okasha, 2015; Luque, 2017; Okasha & Martens, 2016; Rubin, 2018). Within this literature, one key issue is whether HR explains and, if so, how. Our goal in this paper is to reinforce the case for the considerable explanatory value the general regression version of HR (Gardner et al., 2011; Queller, 1992) has for explaining the evolution of social traits. In particular, we argue that we can view this version of HR as either providing an ‘explanation by constraint’ (Lange, 2017, 2018b, see also; 2013) or a non-causal counterfactual explanation (Jansson & Saatsi, 2019; Reutlinger 2016, 2018; Saatsi, 2018b; Woodward, 2018). These discussions often use examples from physics. Our analysis broadens the debate by considering a case in social evolution theory.

The article is structured as follows. In section 2, we introduce the controversy over HR’s explanatoriness. Section 3 reviews one recent solution, namely Birch’s (2017a,b) proposal that HR explains qua being an organizing framework. Although we ultimately disagree with Birch that the rule explains for that reason, we agree that it does not explain by virtue of citing causes. In the following sections, we apply the two accounts of non-causal explanation to HR. Section 4 examines the

∗ The authors are listed in alphabetical order and contributed equally. We would like to thank audiences at the Workshop: Explanation and Understanding in Ghent, the PhD seminar at Erasmus University Rotterdam, the OZSW conference at the University of Twente, the Philosophy and Ethics workshop at Eindhoven University of Technology, as well as participants of the Protective Belt seminar at the London School of Economics for helpful comments and suggestions. We also would like to thank Jonathan Birch for detailed comments on a previous version of the manuscript. Philippe Verreault-Julien acknowledges support from the Fonds de recherche du Québec – Société et culture (FRQSC).

E-mail addresses: koliofotis@esphil.eur.nl, vkoliofot@hotmail.com (V. Koliofotis), p.verreault-julien@tue.nl (P. Verreault-Julien).

1 Website: https://pvjulien.net.

https://doi.org/10.1016/j.shpsa.2021.11.008
Received 18 February 2021; Received in revised form 26 October 2021; 0039-3681/© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
2. Does Hamilton’s rule explain?

IFT is an extension of the genetic theory of natural selection. It is typically expressed by a mathematical inequality called HR. The rule states the conditions under which genes for various types of social traits (i.e. social phenotypes) will spread by the evolutionary process of natural selection. HR demonstrates that whether a social trait undergoes selection hinges on the genetic relatedness between those that interact and by the changes in their expected number of offspring.

Depending on the evolutionary problem at hand, different forms of HR can be derived from evolutionary game theory or principles of population genetics. From a philosophical perspective, this is crucial for understanding the current debate. When evolutionary biologists discuss HR, they may have quite different principles in mind. The original formulation of HR states that an altruistic trait will see its frequency increase if and only if relatedness \( r \) times the trait’s benefit \( b \) is greater than the trait’s cost \( c \), or \( rb > c \) (Hamilton, 1964a,b). Hamilton derived his result based on several assumptions such as one-locus population, no epistasis (or dominance), weak selection, and fitness additivity. Following that result, Hamilton and numerous other evolutionary theorists relaxed the original assumptions and examined whether they could recover a similar result with, e.g., class structure, kin competition, or non-additive fitness payoffs (e.g. Hamilton, 1976; Queller, 1985, 1992).

Hence, more than five decades of extensive theoretical research resulted in a variety of formulations of the type \( rb > c \), often coming with different assumptions under which HR applies (e.g. Frank, 1998, 2013; Lehmann & Rousset, 2014; Queller, 1984, 1992, 2011).

According to Birch and Okasha (2015; see also Birch, 2014), depending on the meaning of ‘cost’ and ‘benefit’, it is possible to distinguish among three main versions of HR: general, special, and approximate. A general version (HRG) is derived from the Price equation (Queller, 1992), with the costs and benefits corresponding to partial regression coefficients. A special version (HRS) emanates from evolutionary game theory (Queller, 1984; van Veelen et al., 2012), with costs and benefits reflecting the payoffs of a game matrix. In the approximate formulation (HRA), the costs and benefits of the regression coefficients are approximated by partial derivatives (Taylor & Frank, 1996).

Against this background, both the philosophical and the biological literature have mainly focused on the general version of the rule derived from the Price equation (Price 1970, 1972). As a mathematical model, the Price equation can be applied to different kinds of evolutionary processes, both biological and non-biological. In biological applications, it is used to provide a representation of the process of natural selection by the covariance between fitness and the trait of interest (e.g. Frank, 2012; see Luque, 2017, for a review of extensions and applications).

According to the Price equation, for a specific trait \( g \), the average change in the value\(^2\) of the trait \( \Delta g \) is

\[
\Delta g = \frac{1}{m} \left[ \text{cov}(w, \Delta g) + E[w, \Delta g] \right]
\]

where \( w \) is individual fitness or reproductive value and overbars refer to population averages. Derivations of HRG often focus on the selection component of the Price equation and, therefore, we can omit the second term of the Price equation.\(^3\) What remains captures the action of natural selection, i.e. the change in the genetic value of the social trait.

\[\text{cov}(w, g) = m \beta_{fg} \quad \text{and} \quad E[w, g] = m \beta_{fg} \rho \]
teractions where the combined effect produced by the behaviour of two or more organisms is higher to the sum of their separate fitness effects), attributing causal meaning to HRG requires particular — and controversial — assumptions about environmental constancy.

The second strategy is more subtle and involves showing that HRG explains, but not because it cites the causes of social evolution. Birch (2017a,b) argues that HRG explains because it is an organizing framework. HRG does not have the same explanatory role as more detailed evolutionary models. Instead, it facilitates evolutionary research by organizing causal explanations into meaningful classes. According to Birch, HRG allows the interpretation, classification, and comparison of evolutionary models that make substantial assumptions about social evolution.

As Birch (2017b, p. 50) observes, this is not a totally novel suggestion. It is common for evolutionary theorists to classify and interpret the results of evolutionary models within the framework provided by IFT (e.g. West et al., 2011). Sometimes evolutionary theorists aim for generality, whilst in many cases they focus on applications to particular problems (e.g. Gardner et al., 2011; Queller, 1992). A general formulation of HR that does not depend on restrictive assumptions is the appropriate tool for conceptual analysis. According to Gardner and his colleagues, HRG can be applied …

[... ] as a conceptual aid for understanding the results delivered by these analyses [mechanistic, dynamically sufficient models], and connecting these with the results of other models analyzed using different methods, as it provides a general framework and universal language for social-evolution theory (Frank, 1998; Gardner et al., 2007; Rousset, 2004; Taylor & Frank, 1996). (Gardner et al., 2011, p. 1037)

Although we think these two strategies may be correct, we also believe another strategy is worth pursuing for two reasons. First, as Birch and Okasha (2015) acknowledge, whether HRG can receive a causal interpretation is a point of contention. It is especially problematic in the case of non-additive effects (Okasha & Martens, 2016; van Veelen, Allen et al., 2017). That puts into question the merits of pursuing that strategy. Second, even if grant that HRG is an organizing framework, we do not believe that it explains for that reason. This is the subject of the next section. In sections 4 and 5 we identify alternative grounds for HRG’s non-causal explanatoriness.

3. The organizing framework defence

Birch (2017a,b; see also 2014) holds that HRG is an organizing framework that provides a classificatory scheme ‘that allows to interpret all models using the $rb > c$ inequality. As such, HRG may classify as similar particular models that cite categorically different mechanisms. For instance, mechanisms of kin discrimination (via environmental or genetic cues) and limited dispersal (population viscosity) have in common that they promote the evolution of social traits by generating a sufficiently high degree of relatedness between interacting organisms. However, this same result is achieved through distinct causal routes. According to Birch (2017b), other models instead emphasize that some traits have a negative cost ($—c$), thus downplaying the need for a positive $rb$; social traits may evolve due to their direct fitness effects. Similarly, various causal routes may lead to negative costs. HRG, Birch says, thus allows classifying the models used to represent all evolutionary target systems.

He gives the following threefold account of an organizing framework. Suppose there is a set of evolutionary models $M$ of ecological scenarios $S$. A model $\Omega$, not itself in $M$, is an organizing framework if:

1. $\Omega$ represents all the target systems represented by the members of $M$, but does so in less detail,
2. $\Omega$ assumes nothing that is not assumed by all members of $M$, and
3. The relations between $\Omega$ and the members of $M$ enable researchers to classify the members of $M$ in an illuminating way (Birch, 2017b, pp. 48–49).

Birch argues that HRG satisfies the first condition because we can view it as an abstract representation of evolutionary change for any genetically inherited trait. It meets the second one because HRG shares some key assumptions with a wide range of evolutionary models that represent change due to natural selection acting in a constant genic environment rather than the change attributed to genetic drift or meiotic drive. Finally, it fulfills the third condition because HRG’s partitioning of evolutionary change into $rb$ and $–c$ components allows to classify explanatory models by how these components interact to bring about evolutionary change. Hence, Birch holds that HRG has the three key features of an organizing framework.

It is true that in its general formulation, HR transcends particular evolutionary models that analyze the genetic evolution of social traits and facilitates contemporary mathematical research on social evolution. If a trait is analyzed correctly, researchers should expect that it would evolve under the conditions specified by HRG. To give an example, over the last decades researchers have advanced a series of strong reciprocity models that explain social behaviour in the laboratory or the field (e.g. Bowles & Gintis, 2004; Gintis, 2000). The claim is that these models provide a distinct and novel account for the evolution of cooperation that is outside the confines of evolutionary explanations provided by IFT. Strong reciprocity refers to a combination of norm-abiding cooperative behaviours and costly punishment of norm violations. However, by distinguishing punishing from cooperative behaviours, West et al. (2011) and Vromen (2012) demonstrate that strong reciprocity could not have evolved unless it provided either direct or indirect fitness benefits. In many cases, punishment behaviour may provide a direct fitness benefit by increasing cooperation within a group. It can be also the case that it may result in indirect fitness returns by decreasing the fitness of competing non-genetic relatives. Hence, this modelling literature does not identify a novel pathway for the evolution of social traits. Instead, it provides an indirect fitness explanation, a direct fitness explanation, or a combination of both (Birch, 2017b).

Before going further, we should say we agree with Birch on two points. First, that HRG can serve as an organizing framework. Birch’s case is convincing and consistent with our reading of the practitioners. Second, we also believe that HRG explains and, in particular, that it provides a non-causal explanation. However, pace Birch, we do not think it explains qua organizing framework. Importantly, these two views can coexist: HRG can be an organizing framework and explain for a different reason than being an organizing framework. In other words, we contend that HRG explains on grounds other than those identified by Birch. Before looking at them, let us examine his claim.

He provides two main arguments for HRG’s explanatoriness. The first consists in inferring that it explains on the basis that it affords understanding. He says that HRG “would be non-explanatory only if it added nothing of value to our understanding of social evolution” (Birch, 2017b, p. 69). According to him, HRG is ‘unifying’ in a way that existing philosophical theories of explanation do not capture (e.g. Birch, 2017b, p. 68, fn. 4). In particular, Birch notes that HRG does not unify in Kitcher’s (1981, 1989) sense of providing argument patterns.

Even though we are sympathetic to this line of argument, whether having understanding implies having an explanation (i.e., explanation is

---

5 Levin & Grafen (2019) explicitly endorse the organizing framework interpretation, but they do not discuss it in the context of explanatory considerations.

6 Birch (2017b, p. 76) also suggests that under a liberal interpretation of ‘causal explanation’ we could say that HRG explains causally. We prefer to stick with the non-causal label since the liberal interpretation involves a significant departure from standard accounts. See also Andersen (2018) for a discussion of broad and narrow construals of causal explanations.
necessary for understanding) is a matter of dispute (e.g. Dellsén, 2020; Gijsbers, 2013; Lipton, 2009; Rice, 2016; Verreault-Julien, 2019b). There are two reasons to question that connection between understanding and explanation. First, it is plausible that we can obtain explanatory understanding without having an (actual) explanation. Some have argued that so-called how-possibly explanations, viz. explanations that do not actually explain phenomena, may afford understanding (e.g. Reutlinger et al., 2018; Rohwer & Rice, 2013; Ylikoski & Aydinonat, 2014). While (actual) explanations may be an important source of understanding, that literature suggests it is not the only route. Likewise, Lipton submits that “we may unify the phenomena (and so improve our understanding of them) by constructing schemes of classification that do not in themselves provide explanations [...]” (2009, p. 54). According to him, Kuhnian exemplars are one plausible mechanism through which we can acquire that knowledge of unification without explanation.

Second, it is not clear whether the sort of understanding HRG affords qua organizing framework is properly explanatory. Gijsbers (2013, see also 2014) proposes a distinction between two sorts of understanding, viz. explanation-understanding and unification-understanding. According to Gijsbers, only the former is explanatory in virtue of identifying what he calls “vertical connections” between phenomena. By contrast, one gains unification-understanding by knowing about the “horizontal connections of kinship” between phenomena. Knowing horizontal connections allows one to see the significant similarities between phenomena; it unifies them. Gijsbers claims that this form of unification affords understanding, yet does not explain. More generally, the worry is whether or not all understanding is of the explanatory sort. In particular, some have argued for an objectual (vs explanatory) notion of understanding (e.g. Dellsén, 2020; Elgin, 2017; Kelp, 2015). Therefore, if we cannot dismiss the possibility that we gain understanding without explanation and if understanding is not necessarily of the explanatory sort, then we cannot straightforwardly infer from the benefit of understanding to the conclusion that HRG actually explains. These debates indicate that whether HRG affords understanding is not, by itself, sufficient to warrant the conclusion that it explains. As a result, we need to look in more detail at the other reasons Birch provides.

This is what the second argument for HRG’s explanatoriness does. Let us then assume that it is correct to infer explanatoriness from the presence of understanding. Why does HRG afford understanding in the first place? Of the three conditions Birch states for a model Ω to serve as an organizing framework, we take it that the third one is the most important with respect to its explanatoriness. This is because it would be possible to build an arbitrary framework Ω that satisfies conditions (i) and (ii), but without any explanatory value. What matters is that Ω’s classification of models M explains. For Birch, HRG explains because it “generates understanding of causes by providing a framework for comparing, classifying, and interpreting such explanations” (2017b, p. 76). Elsewhere, he also says that “[t]he insight embodied by HRG is that every adequate causal explanation of positive change can be placed somewhere in this space” (2017a, p. 4). In short, the point is that HRG provides an illuminating way to classify evolutionary explanations and since that classification affords understanding, then it also explains. For instance, HRG may help to see that different causal mechanisms increase rb and therefore have the same effects. Or, if we observe that a given trait is costly, then we can infer that it must be explained by its indirect fitness effects rb.

We agree that HRG provides an illuminating classification of evolutionary models. However, explanations are typically judged by their capacity to increase our understanding of empirical phenomena of interest. According to Birch, what HRG helps us to understand are the various evolutionary explanations it organizes. It provides insights on evolutionary theory by organizing explanations on the basis of their commitment to the sign of r, b, and c (Birch, 2017b, p. 46ff.). Indirect fitness explanations, for instance, “rely on there being a mechanism that explains the systematic tendency for the benefits caused by the expression of genes for altruism to fall differentially on other bearers of those genes […]” (Birch, 2017b, p. 51). Limited dispersal can be such a mechanism that explains why the benefits caused by altruistic genes are differentially directed to those that have those genes. HRG shows that all these indirect fitness explanations share in common rb > 0 and c ≥ 0.

The main issue with this argument is that what we appear to understand better with HRG is not empirical phenomena, but social evolution theory. Put differently, HRG qua organizing framework affords theoretical, not explanatory understanding. De Regt’s (2017, see also 2009; De Regt & Dieks, 2005) distinction between understanding a theory and understanding phenomena is helpful to make this point plain. According to him, a “phenomenon P is understood scientifically if and only if there is an explanation of P that is based on an intelligible theory T and conforms to the basic epistemic values of empirical adequacy and internal consistency” (De Regt, 2017, p. 92). Intelligibility is a pragmatic notion and refers to a cluster of virtues that theories possess. It is thus context-dependent in that different epistemic communities may assess the intelligibility of given theories differently.

What is important for our purposes is that although having an intelligible theory is a necessary condition for understanding phenomena, it is not sufficient. Scientists also need to use the theory to construct (correct) explanations. We take it that what HRG does qua organizing framework is to improve the intelligibility of social evolution theory. As such, it facilitates the construction of explanations. But theories and explanations are not the same things and facilitating the construction of explanations is not the same as explaining. The explanations explain, not the intelligible theory. In De Regt’s terms, while understanding a theory is a condition for understanding phenomena, understanding phenomena is the “product of explanations” (De Regt, 2017, p. 92). Explanations, in turn, may have to satisfy different conditions. For instance, proponents of causal explanations hold that all explanations must cite causes. Others are pluralist. De Regt himself endorses a rather minimal set of conditions by saying that adequate explanations only need to be empirically adequate and internally consistent. If this is correct, then what would explain are the explanations we can construct using HRG, e.g. indirect fitness explanations. So HRG could be instrumental in building explanations, but in and of itself would not explain.

At this point, one could retort that if an organizing framework improves the intelligibility of social evolution theory, then does it not also help to understand phenomena? For if understanding a theory is necessary for understanding phenomena, then surely improving our grasp of social evolution theory is instrumental to our understanding of social evolution phenomena. We have nothing against this conclusion. It is plausible to suggest that the degree of understanding phenomena depends on a theory’s intelligibility (see De Regt, 2017, p. 100, fn. 10). However, our main contention stands, namely that despite HRG’s positive contribution to our understanding of social evolution theory, that does not make it an explanation of social evolution phenomena. We thus believe that if HRG explains, it must be for reasons other than being an organizing framework. In the section that follows, we provide these reasons.

4. An explanation by constraint

The dominant account of explanation is the causal one (Reutlinger, 2017; see also Strevens, 2008; Woodward, 2003). It is thus not surprising that evolutionary researchers have aimed at showing how HRG may

---

7 This also suggests we could interpret the disagreement between the critics and supporters of HRG as one concerning the explanatory status of HRG, i.e. whether it provides an actual or a possible explanation (see, e.g., Verreault-Julien, 2019a).

8 Yet other ways to frame the problem would be to say that HRG affords understanding-with (Strevens, 2013). See Newman (2017) for an account of theoretical understanding.
identify the causes of social behaviour (e.g. Okasha & Martens, 2016). However, a growing literature has challenged that predominance by investigating putative cases of non-causal explanations, in particular mathematical explanations (Baron, 2019; Baron et al., 2017; Jansson & Saatsi, 2019; Lange, 2013; Pincock, 2015; Raz, 2017; Reutlinger, 2016; Saatsi, 2012).\(^9\) Basically, what these authors argue is that there are bona fide explanations that explain despite them not citing causes.

In what follows, we explore two types of accounts that illuminate the debate over HRG’s explanatoriness. The first type are ‘explanations by constraint’ (Lange 2013, 2017, 2018b). This will be the object of this section. The second type is what we call the non-causal counterfactual account (e.g. Jansson & Saatsi, 2019; Reutlinger, 2016, 2018; Woodward, 2018), which we discuss in section 5. We use both of these accounts for two main reasons. First, we do not want to take a stand concerning whether there is a single best account of non-causal explanations or whether one is superior to the other. Second, if we can substantiate that there are plausible readings of HRG as either type of non-causal explanation, then this strengthens our claim that we can, indeed, interpret HRG as providing a non-causal explanation.\(^10\)

Explanations by constraint show that the explanandum is due to facts that are more necessary than ordinary causal generalizations. Consider the following simple example.

The fact that 23 cannot be divided evenly by 3 explains why Mother fails every time she tries to distribute exactly 23 strawberries evenly among her 3 children without cutting any (strawberries—or children!) (Lange, 2017, p. 6).

Lange argues that what explains Mother’s failure to divide evenly the strawberries is not the particular physical facts of the situation. Rather, it is the modally stronger mathematical facts. It is just impossible to divide evenly 23 (whole) units of anything. Although we could make the division possible by altering the physical conditions (e.g., by distributing 24 strawberries), the facts that prevent (or would allow) the division are mathematical, not physical.

Lange holds that many scientific explanations function similarly by appealing to ‘constraints’. He says that these explanations …

[...] work not by describing the world’s network of causal relations in particular, but rather by describing the framework that any physical system (whether or not it figures in causal relations) must inhabit, where this variety of necessity is stronger than the necessity possessed by the ordinary laws of nature (Lange, 2017, p. 44, see also Lange, 2015).

For Lange, explanations by constraint are non-causal since what explains are not causal facts, but facts — constraints — of a stronger modality. A crucial aspect of Lange’s account is the idea that there are different levels of necessity (see also Lange, 2009). What matters is that the constraints’ modal force transcends that of weaker causal facts. Constraints, then, just are facts that manifest a stronger necessity than the particular causal details of a situation.

For example, in the case of Mother and her strawberries, we could explain her failure by citing her repeated and unsuccessful attempts at evenly sharing the 23 strawberries with her 3 children. The causal process underlying the distribution of the strawberries does not bring about even shares between the children. But this causal process is modally weaker than the mathematical facts. Thus, these facts constrain the causal details. Regardless of how Mother would try to distribute the strawberries, she would fail. Hence, if the explanandum occurred it is because of these constraints, not because of the causal details.

We believe that HRG can be interpreted as explaining in virtue of being a constraint. HRG transcends causal explanations because it delineates the biological (causal) space of possibility. It provides a formal and conceptual framework that all causal explanations need to satisfy. Let us illustrate this with the following why-question: Why has altruism never evolved without positive relatedness?\(^11\) According to the standard definition of evolutionary (or biological) altruism, it is a behaviour that results in a decrease in the actor’s lifetime fitness and in an increase of the recipient’s lifetime fitness. We could explain this fact by providing the causal details of the different situations, for instance by showing that kin discrimination was absent and thus could not result in positive assortment. Or, we could show that in another case the lack of positive assortment was due to an absence of phenotypic markers. By doing this we would answer the why-question by indicating how the organisms were not positively related in virtue of the causal details. In other words, we would provide a causal explanation. But we can also explain why altruism has never evolved without positive relatedness by appealing to HRG. As we have seen in section 2, if \(\mathbf{AG}\) denotes the average change in the frequency of a gene between the ancestral and descendant populations, natural selection favours genetically transmitted social traits when:

\[
\Delta g > 0 \text{ iff } rb - c > 0 \text{ given that } var(g) \neq 0
\]

What this implies is that regardless of the social trait under examination, how the population is structured, or how the trait relates to costs and benefits, it was selected because \(rb > c\). Benefits \(b\), costs \(c\), and genetic relatedness \(r\) are all regression coefficients and thus might not represent causal processes (see section 5 below). Although \(r\) can be negative in the case of spiteful behaviour, altruism cannot evolve if relatedness is equal to zero. For if \(r\) is zero, then the indirect fitness term \(rb\) equals zero too. This leads to the mathematically impossible result \(-c = \beta_{mg} > 0\) (a negative number cannot be higher than zero). Formally, HRG requires a positive \(r\) for altruism to evolve.

Therefore, altruism could not have evolved in the absence of positive relatedness simply because it did not fulfill the conditions stated by HRG. Crucially, HRG holds no matter the causal details. For instance, if altruism evolved in a population, HRG would have still held if different causal processes had been responsible for the evolution of altruism. As long as behaviour comes at a net cost for the actor and net benefit for the recipient, relatedness needs to be positive. The particular details giving costs, benefits, and relatedness their values are irrelevant. In that sense, HRG acts as a constraint since it is modally stronger than the causal generalizations it applies to; it constrains potential causal explanations of the trait’s evolutionary change. Any causal explanation of social traits will be consistent with the \(rb - c > 0\) inequality. What practitioners say supports this interpretation.

It is simply incorrect to claim that Hamilton’s rule requires restrictive assumptions or that it almost never holds (e.g. van Veelen, 2009; Nowak et al., 2010). [...] [HRG] applies to any scatter of genetic and fitness data, irrespective of the underlying causes of this variation. It makes clear that, whatever the relationship between trait and reproductive success, the resulting action of natural selection can be decomposed

\(^{9}\) Here, ‘mathematical explanations’ means scientific explanations of empirical phenomena that explain in virtue of — non-causal — mathematical facts. We are not here referring to explanations within mathematics (see Mancosu, 2001; Steiner, 1978).

\(^{10}\) Whether explanations by constraint and non-causal counterfactual explanations are different species is contentious. For instance, Saatsi (2018a) considers that explanations by constraint can be subsumed under the more general counterfactual umbrella. Our argument does not hinge on that debate since our goal is to show that HRG explains according to either one of these accounts.

\(^{11}\) It would also be possible to formulate the why-question in a slightly different way: Why is it impossible for altruism to evolve without positive relatedness? We chose the former formulation because the explanandum is clearly empirical whereas in the latter it would be a modal fact, which might — incorrectly — suggest that HRG does not explain biological phenomena. That said, either formulation is amenable to an explanation, so our argument does not hinge on that (Lange, 2018b).
into cost, benefit and relatedness terms (Gardner et al., 2011, p. 1037, our emphasis).

Gardner et al. make two claims of particular interest. Firstly, they reject the interpretation according to which HRG rarely holds (cf. Nowak et al., 2010). In other words, HRG actually explains and applies to any data. Secondly, they assert that HRG holds regardless of the particular causal details. We take this particular point to be crucial for our claim that HRG explains qua constraint. HRG's explanatory contribution does not lie in its specification of the particular causes of the phenomena it purports to explain. Instead, it applies to all fitness and genetic data, irrespective of the causal details. No matter how one wants to explain the evolution of particular social traits, the relationship between benefits, relatedness, and costs will be related as per HRG states. In fact, not only HRG does not specify how individuals are causally related in a given population, but for HRG these causal details are irrelevant. In a given explanatory context, HRG implies there is a degree of necessity that goes beyond what ordinary causal explanations would tell us.

One could now ask whether explaining qua constraint is really different than explaining qua organizing framework since an organizing framework arguably sets boundaries to causal explanations. Our reply is twofold. First, the explanation by constraint account emphasizes the actual feature that is responsible for HRG's explanatoriness. Being an organizing framework is accidental to whether a principle like HRG explains. Some organizing frameworks explain because they constrain; some constraints are not organizing frameworks. We identify what is actually responsible for HRG's explanatoriness, namely that it acts as a constraint. Organizing frameworks and constraints are not coextensive. For example, a classification of mammals arguably satisfies all of Birch's (2017b) criteria for an organizing framework, but does not explain. Conversely, some laws of nature (e.g. Newton's second law, see Lange, 2017) may act as constraints without being an organizing framework in the sense of Birch (2017b). Second, as we have said above, it is unclear how HRG qua organizing framework explains phenomena since its main function is to classify explanations. Whereas HRG qua constraint tells us why altruism evolved (because of positive relatedness), an organizing framework does not provide such answers. Therefore, we believe that the explanation by constraint account better locates the source of HRG's explanatoriness.

It may also be possible to object that there is an important disanalogy between HRG and the case of Mother and her strawberries. Whereas the latter invokes mathematical facts, HRG describes the conditions under which natural selection will favour social traits. It is relatively uncontroversial that mathematical facts are modally stronger than causal ones, but HRG does not seem to fit squarely in either category.12 Although delving into the details of Lange's (see 2009, 2017) account of degrees of necessity could be interesting, the only important thing for our purposes is that HRG is modally stronger. Within Lange's framework, HRG's modal status is analogous to that of symmetry principles and conservation laws. That particular gravitational or electrical interactions conserve energy can be the result of, respectively, gravitational or electrical force. We could thus explain energy conservation by appealing to these forces. However, Lange argues, this amounts to treating conservation of energy as a coincidence. Instead, energy is conserved because the law of conservation of energy requires force laws to conserve energy. This is no coincidence; conservation laws constrain force laws. In turn, symmetry principles constrain conservation laws. For example, the law of conservation of energy follows from time translation symmetry. Even though conservation laws are not mathematical truths, they are more necessary than the force laws they constrain.

Current interpretations from practitioners and philosophers alike support this weak claim of relative modal strength for HRG. Indeed, they often express its contribution with a biconditional relation between the explanandum and explanans. Birch's (2017a,b) characterization is explicit: \( \Delta g > 0 \iff rb > c \). Likewise for Gardner & West (2014; 1, our emphasis; see also Gardner 2015) who write that \( \lbrack \text{HRG}\rbrack \) states that any trait—altruistic or otherwise—will be favoured by natural selection if and only if the sum of its direct and indirect fitness effects exceeds zero. Causal explanations typically do not identify necessary conditions. There is a multitude of causes that may be responsible for the evolution of altruistic traits: limited dispersal (i.e. population viscosity) keeps relatives together or kin discrimination based on environmental or genetic cues typically generate a high degree of genetic relatedness among interacting individuals. Both, therefore, favour the evolution of altruistic cooperation. What causes the costs, benefits, and relatedness to have particular values may be different across populations. The key point is that regardless of the causes at play, the relationship between \( r, b, \) and \( c \) will hold. In that sense, it is more necessary than the particular causes.

Furthermore, both champions (e.g. Frank, 2012) and critics (Allen et al., 2013; Nowak & McAvoy, 2017; van Veelen & Garcia, 2012) alike point out that HRG is akin to a tautology. If HRG is indeed modal stronger than the causal details it applies to, then we should expect HRG to appear as such. Moreover, we should also expect that to refute it would require meeting a different set of conditions. Whereas we can see the inadequacy of a causal explanation by the lack of a causal relationship between the explanans and the explanandum, this would not falsify the constraint. The constraint holds regardless of the causal details. This is why we take HRG's apparent tautological nature as evidence for our claim that HRG has a stronger degree of necessity than particular causal explanations of the same explanandum.13 HRG thus lays out a mathematical inequality that all causal explanations need to satisfy. Regardless of how one wants to causally explain particular evolutionary phenomena, the relationship between benefits, relatedness, and costs will abide by HRG.

5. A non-causal counterfactual explanation

In the previous section, we argued that it is plausible to interpret HRG as providing an explanation by constraint. We believe there is also a second, distinct, non-causal reading of HRG's explanatoriness following the non-causal counterfactual account of explanation (Janssøn & Saatsi, 2019; Reutlinger 2016, 2018; Saatsi, 2018b; Woodward, 2018).

The non-causal counterfactual account extends Woodward's (2003) interventionist theory of causal explanation. According to it, causal explanations allow us to answer 'what-if-things-had-been-different' questions (what-if questions henceforth). They tell us what would happen to \( Y \) under a hypothetical intervention on \( X \). The basic idea behind the non-causal counterfactual account is that although explanations always provide information about counterfactual dependence, that dependence need not be causal in the interventionist sense.

To illustrate, let us consider again the simple example above of Mother and her strawberries.14 Janssøn and Saatsi (2019) and Woodward (see also Woodward, 2018) argue that we can interpret it along counterfactual lines. The counterfactual ‘Had Mother had 24 strawberries, then she would have been able to divide them evenly between her 3 children’ is true (cf. Lange, 2017, pp. 19–20). We can evaluate the truth of that counterfactual because we know that Mother's failure — or capacity — to divide evenly the strawberries depends on the number of strawberries she has. The facts that support this counterfactual rely on mathematics. They

12 Here we sidestep the debate over whether the principle of natural selection is a priori or not (Elgin & Elliott, 2015; Lange & Rosenberg, 2011; Sober, 2011).
13 To be clear, we do not believe HRG is a tautology. Rather, it just looks like one. We discuss this issue in more detail in section 5.
14 This does not imply that the two types of non-causal explanations are similar. Even though both accounts would consider this example to be of the non-causal sort, they reach that judgment for different reasons (see Reutlinger & Saatsi, 2018). We remain agnostic concerning the relationship between these two types.
are not causal. We know that if Mother's number of strawberries had been adequately different, then she would have been able to divide them evenly. Crucially, the counterfactual does not depend on the particular physical details of the situation, but rather on mathematical facts.

Does HRG explain in virtue of citing non-causal difference-making relations between $\Delta g$ and $r$, $b$, and $c$? We believe so. But what makes a counterfactual explanation causal? Indeed, to show that HRG provides a non-causal counterfactual explanation, we can first ask whether it satisfies the criteria for causal explanation. If it does, this would indicate that our proposal is misguided. But if it does not, then this opens the door to a non-causal interpretation.

According to Woodward (2018, p. 122, see also Woodward, 2003), causal explanations have the following three features:

1. They provide answers to what-if questions that tell how $Y$ changes.
2. Under possible physical interventions on one or more ($X_1, \ldots, X_n$), and
3. The relationship between $X_1, \ldots, X_n$ and $Y$ is empirically (not conceptually) invariant under a range of interventions on $X_1, \ldots, X_n$ and background conditions.

Woodward suggests that we may have a non-causal explanation when we relax one or more of these conditions. The most important option involves keeping feature (1), but relaxing (2) or (3). In other words, non-causal explanations may allow to answer what-if questions, but by appealing to non-causal counterfactual dependence. For instance, the case of Mother and her strawberries would violate (3) — the dependence is mathematical — and would satisfy (1) and (2) — we can intervene on the number of strawberries and it allows us to answer what-if questions. Let us look at these conditions in turn.

Does HRG provide answers to what-if questions? It seems to be a straightforward case of counterfactual dependence. The rule says that a change in the frequency of a trait depends on genetic relatedness as well as fitness costs and benefits. Since HRG is derived from substituting the Price equation into a regression equation, $c$ and $b$ are partial regression coefficients that relate changes between variables: had the value of $b$ been different, then the value of $\Delta g$ would have been different. HRG thus allows us to answer what-if questions because of the counterfactual dependence relation between the variables. Should $r$, $b$, and $c$ be interpreted as causes of evolutionary change? In the case of HRG, we think it is more appropriate to consider that the counterfactual dependence is non-causal. This is because statistical association does not necessarily imply causation and, therefore, HRG might not express a causal relationship. The second step, then, is to look at whether or not the relation of counterfactual dependence that HRG expresses is causal.

Woodward's second criterion requires that it is possible to physically intervene on the explanatory variables.13 For instance, one may explain the occurrence of bad weather by showing that had the atmospheric pressure been higher following an intervention, then the weather would have been fair. This relationship between atmospheric pressure and the weather is invariant under a range of interventions. However, intervening on the barometer needle would not change the weather. That we could hypothetically intervene on the atmospheric pressure and change the weather, but not on the barometer needle, indicates that the two are not causally related — only atmospheric pressure is a cause of the weather.

The intervention criterion is at the centre stage of the discussion on HRG's causal interpretation. In short, the problem is to see whether or not HRG specifies what would happen upon hypothetical physical interventions (e.g. Birch, 2017b, p. 72). As we indicated in section 2, the consensus so far in the literature is that there is no straightforward interpretation of HRG that satisfies the intervention criterion (2) (Birch, 2017b; Nowak & McAvoy, 2017; Okasha, 2016; Okasha & Martens, 2016). In a recent analysis of HRG, Okasha and Martens (2016) consider a hypothetical experiment — an intervention — that randomly draws a selfish type from the population and switches it into an altruistic type. Then, they compare the results to the partial regression definitions of costs and benefits. They show that these definitions plausibly have a causal meaning in the case of additive payoffs. This is because the expected causal effect of the intervention corresponds to the values of the partial regression coefficients of HRG. In other words, we can interpret the regression equation as representing how a change to an individual's genetic value makes a (causal) difference to that individual's fitness. However, in the more general case of non-additive pairwise interactions, Okasha and Martens argue that costs and benefits in HRG do not have a clear-cut causal meaning if the selfish types are picked at random from the population. This is because the $-c$ variable would not represent the per capita causal effect of an intervention. With non-additivity, $-c$ and the expected causal effect of the intervention "will almost always differ in magnitude, and may differ in sign" (Okasha & Martens, 2016, p. 5).

Essentially, the problem is not that we cannot physically intervene on the genetic value or that interventions cannot satisfy Woodward's (2003, p. 98f.) criteria for possible interventions. Rather, the problem is that if the intervention is physical, then HRG cannot provide correct answers to what-if questions when non-additivity holds.

Okasha and Martens (2016) propose one way of rescuing the causal interpretation. They show that the cost in HRG corresponds to the expected outcome of the hypothetical experiment only if the selfish types chosen to be experimentally manipulated are drawn from a specific cohort that meets a Fisherian condition of environmental constancy. This is not an innocuous assumption. As they observe, this understanding of environmental constancy is particular to a simple evolutionary model of pairwise interaction with synergistic payoffs used to assess the cost of the social trait and a corresponding measure of assortment that lacks independent justification. But even if we grant that HRG can receive a causal interpretation when additive effects hold, that actual social interactions are non-additive is the more realistic assumption (Grafen, 2006, p. 543). And since the point of deriving HRG in the first place was to present a general and mathematically valid expression of HR (Queller, 1992), limiting the scope of HRG to additive effects would require restricting its intended domain of application. Hence, a causal interpretation of HRG along interventionist lines comes at a high cost.

If HRG cannot receive a causal interpretation in realistic evolutionary scenarios, then where does that leave us? HRG is not invariant under physical interventions because these may fail to bring about the expected changes in the case of non-additivity. Nevertheless, HRG exhibits a statistical relationship: it relates changes to the value of $\Delta g$ to changes in the values of $r$, $b$, or $c$. Moreover, that relationship is invariant in that it holds for all changes to the value of the variables. Not all statistical relationships are invariant in that sense. For instance, a statistical generalization represented in economics by the Phillips curve refers to the stable and inverse relationship between unemployment and inflation (or rate of nominal wage change). Due to stagflation in the 1970s, economists learned that this relationship was less invariant than they thought. In contrast, HRG is highly invariant and this is one reason why it is so valuable as a statistical description of the conditions under which social traits may evolve.14 For example, suppose that $r$ has a positive value and that $b$ and $c$ are such that $rb > c$. Under these assumptions, altruism may evolve, viz. $\Delta g$ would be positive. However, had $r$ been negative, we can infer that $rb > c$ could not have been the case. The upshot is that, understood statistically, HRG allows us to answer what-if questions concerning the different values $\Delta g$ would take if $r$, $b$, or $c$ had been different. Therefore, if HRG explains, it is not in virtue of satisfying

---

13 What a 'possible physical intervention' means has been a subject of debate in the literature (e.g. Reutlinger, 2012). In a nutshell, the problem is that it seems the mere conceptual possibility of an intervention is not stringent enough, but that its actual realization is too demanding. Our discussion does not hinge on a specific solution to this problem.

14 See Woodward (2003, sec. 6.4) for a discussion of degrees of invariance in the context of causal explanation.
criterion (2). And if it does not satisfy that criterion, then the explanation would be non-causal.

Does this mean HRG also fails to meet condition (3), viz. that it holds for purely conceptual or mathematical reasons and not empirical ones? As we noted above, critics find fault with HRG’s apparent tautological character (Allen et al., 2013; Nowak & McAvoy, 2017; van Veelen, Allen et al., 2017; van Veelen & Garcia, 2012). The worry is that HRG is trivially true and thus empirically empty. In particular, the issue is that since HRG does not assume anything about the dynamics that underlie r, b or c, it is not possible to specify how the value of \( \Delta g \) would change over time. We agree with Birch and Birch (2017a,b) that HRG is not in the business of providing particular causal explanations. For that purpose, one could employ models from population genetics or evolutionary game theory (Gardner et al., 2011, p. 1037). Using HRG instead amounts to what Weisberg (2013, sec. 7.1) calls ‘generalized modelling’. HRG’s target is not specific instances of, for example, altruism spreading in a population of eusocial insects. HRG’s target is behavioural patterns that have features shared by all social behaviours, regardless of the detailed life histories of individuals that belong to a particular population. So it is not so much that HRG makes idealizations rather than it abstracts away from the particular details (Rubin, 2018). HRG states that the relationship between r, b, and c is invariant regardless of the causal details; the statistical relationship holds for any population.

That said, we nevertheless believe that HRG is not trivially true in the sense that it expresses an empirical relationship. What is important to observe is that HRG specifies the conditions that need to be satisfied for social traits to evolve. For instance, it states that altruism may only be selected for if there are indirect fitness benefits. Hence, one acceptable answer to the question ‘Why did that altruistic trait fail to spread?’ would be that it had insufficient indirect fitness benefits. This may seem uninformative, but it still provides relevant and true information. And, crucially, that the evolution of an altruistic trait depends on indirect fitness benefits is not a mathematical or conceptual truth. There is thus an important disanalogy between the strawberries case and HRG. Empirical evidence cannot falsify the claim that Mother will never be able to divide evenly 23 by 3. The truth conditions of that claim are mathematical. On the contrary, that natural selection operates via direct or indirect fitness is a hypothesis about the world which empirical evidence can confirm or infirm. One of the common characteristics of inclusive fitness models that provide testable predictions is the partition of fitness into direct or indirect components (e.g. Frank, 1998; Rousset, 2004). This carries over to HRG which partitions fitness using a multivariate regression equation. If empirical research were to demonstrate that indirect fitness (or positive genetic relatedness) is not required for altruism to evolve, then HRG could be rebutted. Even though we may not directly test HRG, the conditions it states for the evolution of social traits can be, and are, the subject of empirical research. Again, there is a fruitful analogy between HRG and conservation laws in physics (see section 4). We treat conservation laws as more fundamental than ordinary force laws. Yet, suppose we were to model a system using the force laws and that empirical observation would establish that the system did not conserve energy (e.g. a perpetual motion machine of the first kind). This would provide evidence against the law of conservation of energy despite the fact that it was not directly used to model the system. And whether systems conserve energy is as much empirical as whether altruism depends on indirect fitness benefits. That a law or principle is considered to be fundamental does not make it immune to empirical evidence.

We have argued that HRG allows answering what-if questions and is empirical, but does not (always) support physical interventions. It thus satisfies the first and third of Woodward’s (2018) criteria for causal explanation, but not the second. This suggests that HRG provides a non-causal counterfactual explanation. But as Woodward observes, accepting non-interventionist counterfactuals creates an issue concerning explanatory relevance. The notion of intervention helps to capture explanatory asymmetry. To use a stock philosophical example, intervening on the position of the sun would make a difference to the length of the shadow cast by the flagpole (Bromberger, 1966). But the sun surely would not move if we were to intervene on the shadow. A cause explains its effect, not the other way around. It is the time-honoured problem of the asymmetry of explanation.

Prima facie, it seems HRG runs into this predicament. HRG expresses correlations and correlations are symmetric. If explanations are asymmetric, how could we explain with HRG? We want to explain \( \Delta g \) by appealing to r, b, and c, but \( \Delta g \) does not explain these variables. However, HRG supports counterfactuals in both ways: had \( \Delta g \) been different, then r would have been different too. How best to capture the asymmetry in non-causal explanations is an ongoing debate (e.g. Craver & Povich, 2017; Khalifa et al., 2021; Lange, 2018a, 2021). Although it can be a problem for particular cases, we do not believe it poses a particular challenge in this case.

First of all, we should note that some non-causal explanations are symmetric in a similar derivational sense (Reutlinger, 2018; Woodward, 2018). Thus, the fact that the counterfactual dependence is symmetric is not sufficient ground to rule out explanatory asymmetry. What it tells us is that counterfactual dependence by itself does not entail explanatory relevance, a conclusion we do not want to resist (see Pincock, 2018). But then, what could be the source of explanatory asymmetry? We are sympathetic to Lange’s (2021) suggestion that there may not be sufficient and necessary conditions for explanatory asymmetry. Hence, the criteria that apply in one case may not be adequate in another. In fact, we believe that in that respect HR is a peculiar case. Initially, Hamilton’s motivation was to find a causal explanation for altruism (Frank, 2013). His original analysis emphasized the causal decomposition of total fitness effects into the relatedness, costs, and benefits of social action. Following Hamilton, many theoretical biologists refined this analysis and developed methods that connect HR to genetics and evolutionary game theory. In this process, Queller (1992) derived a general regression version of the rule that is valid in cases where there is a complicated relationship between genotype, phenotype, and fitness. But, ultimately, what indirectly vindicates HRG is the empirical success of predictions derived from models that apply methods from population genetics and evolutionary game theory. HRG is just a generalization of these models (Queller, 1992, p. 377) and starting from HRG, one can mathematically derive the main result of particular models that make empirical predictions (Gardner et al., 2011). Moreover, it is those models with mechanistic details that provide the direction of explanation. Fitness effects and relatedness explain whether a trait is selected and not vice versa because this is what we model and observe empirically.

Now, would HRG explain in a world with only HRG, but without the other less general versions of the rule and the empirical evidence that supports them? No. But the fact that less general models work gives us reasons to believe that the more general and non-causal one, HRG, also works. So although the justification of HRG comes from more detailed causal models, it does not explain by virtue of specifying causes. HRG would be devoid of any empirical meaning or justification if not for its causal counterparts. This is why, in practice, biologists cannot avoid

---

17 To be clear, the suggestion is not that HRG is a non-causal explanation because it is abstract. We agree with Reutlinger and Andersen (2016) that a high level of abstraction is not a good demarcation criterion between causal and non-causal explanations. For a related discussion, see Bokulich (2014, 2018).

18 One reviewer helpfully remarked that we might argue that HRG does not satisfy Woodward’s third criterion if we restrict ‘empirical relationships’ to those that have weak modal strength in Lange’s sense (2017). This suggests a more nuanced delineation of the empirical vs non-empirical distinction. Although this is an interesting proposal, our main concern here is simply to reject, pace the critics, that HRG holds only for mathematical or conceptual reasons.

19 See Gardner (2020) for a similar argument concerning the Price equation.

20 This is also why we could not simply choose any predictor of fitness and calculate HR, e.g. the moon phase (see Queller, 2011), even though it would lead to a mathematically valid version of the rule. The moon phase is not causally related to fitness and the evolution of social traits.
using less general models based on HRS and HRA (Bourke, 2014; Gardner et al., 2011). But being able to show that the relation described by HRG is invariant also strengthens the justification for the causal models. Models have to explore mechanisms related to one of the two evolutionary pathways that affect inclusive fitness and select for social traits. For instance, the invariance makes it extremely unlikely that we would ever find altruism without mechanisms that generate sufficiently high indirect fitness benefits. This is true regardless of the particular causal details. And that we do not find altruism without those mechanisms reinforces the case for HRG. There is thus a sort of feedback relationship between the causal and non-causal versions of HR. In that sense, HRG and the more detailed causal models complement each other (Andersen, 2018).

To sum up, HRG would satisfy criteria (1) — answering what-if questions — and (3) — invariant for empirical reasons —, but not (2). If HRG explains and does so in virtue of identifying a relation of counterfactual dependence, but does not satisfy (2), then it provides a non-causal counterfactual explanation.21 Of course, this assumes that intervention is necessary for causation. This assumption is sometimes disputed (Reutlinger, 2012; Stevens, 2008). We have focused on the interventionist framework because it is the one HRG has been assessed against in the literature. Other theories of causal explanation may thus reach a different verdict. Our goal is not to settle this question. If it turns out that HRG can receive a plausible causal interpretation, so be it. Our aim was more modest, viz. to show that HRG can explain despite a lack of causal interpretation.

6. Conclusion

We have argued that even though HRG may serve as an organizing framework, it does not explain for that reason. Since we share the idea that HRG does not explain in virtue of citing causes, we have proposed to interpret HRG as providing a non-causal explanation according to two other leading accounts. More precisely, we maintain that HRG can either be viewed as providing an explanation by constraint or a non-causal counterfactual explanation. These accounts better identify the reasons why HRG explains the evolution of social traits.

References

Abbot, P., et al. (2011). Inclusive fitness theory and eusociality. Nature, 471(7339), E1-E4. https://doi.org/10.1038/nature09831
Allen, B., Newak, M. A., & Wilson, E. O. (2013). Limitations of inclusive fitness. Proceedings of the National Academy of Sciences, 110(50), 20135–20139. https://doi.org/10.1073/pnas.1317588110
Andersen, H. (2018). Complements, not competitors: Causal and mathematical explanations. The British Journal for the Philosophy of Science, 69(2), 485–508. https://doi.org/10.1093/bjps/axw023
Baron, S. (2019). Mathematical explanation by law. The British Journal for the Philosophy of Science, 70(3), 683–717. https://doi.org/10.1093/bjps/axz062
Baron, S., Glynn, M., & Ripley, D. (2017). How mathematics can make a difference. Philosophers’ Imprint, 17(3), 1–19.
Birch, J. (2014). Hamilton’s rule and its discontent. The British Journal for the Philosophy of Science, 65(2), 381–411. https://doi.org/10.1093/bjps/axn016
Birch, J. (2017a). The inclusive fitness controversy: Finding a way forward. Royal Society Open Science, 4(7), 170335. https://doi.org/10.1098/rsos.170335
Birch, J. (2017b). The philosophy of social evolution. Oxford: Oxford University Press.
Birch, J., & Okasha, S. (2015). Kin selection and its critics. BioScience, 65(1), 22–32. https://doi.org/10.1093/biosci/biu096
Bokulich, A. (2011). How scientific models can explain. Synthese, 180(1), 33–45.
Bokulich, A. (2014). How the tiger bash got its stripes: ‘How possibly’ vs. ‘how actually’ model explanations. The Monist, 97(3), 321–338. https://doi.org/10.5840/themonts2014-97221
Bokulich, A. (2018). Searching for non-causal explanations in a sea of causes. In A. Reutlinger & J. Saatsi (Eds.), Explanation beyond causation: Philosophical perspectives on non-causal explanations (pp. 141–163). Oxford: Oxford University Press.
Bourke, A. F. G. (2014). Hamilton’s rule and the causes of social evolution. Phil. Trans. R. Soc. B, 369(1642), 20130362. https://doi.org/10.1098/rstb.2013.0362

21 Cases with a similar structure have been discussed in the literature (e.g. Bokulich, 2011, 2018; Reutlinger, 2016; Rice, 2015; Saatsi, 2018c; Saatsi & Pexton, 2013).
