European Journal for Philosophy of Science (2026) 16:11
https://doi.org/10.1007/s13194-026-00720-5

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

Self-defeating and self-fulfilling reactivity

Yotam Harel[1]

Received: 12 September 2024 / Accepted: 10 January 2026

Abstract
Theory-deduced predictions might change agents' beliefs, and thus also agents' behavior. Since agents react to their beliefs by modifying their behavior to obtain their goals, they might react to a belief inspired by a theory-deduced prediction by modifying their behavior to obtain their goals, and this may have implications for the theory and its predictive success. In this paper, I first theorize this phenomenon. I disqualify past formulations of so-called reflexive predictions and advocate my account of self-defeating and self-fulfilling reactivity. I then examine the implications of three kinds of self-defeating reactivity, weak, strong, and vicious, for predicting. I conclude that self-defeating reactivity makes it impossible to predict, at least in some cases. Finally, I rethink whether self-defeating and self-fulfilling reactivity is exclusive to the human/social sciences or to states of affairs where human beings/social actors are involved. Here, I conclude that while reactivity is not exclusive to the human/social sciences, it is exclusive to cases where agents are involved. Thus, it is exclusive to cases where human beings/social actors are involved only *de facto*.

Keywords Reactivity · Reflexivity · Predictions · The exclusivity theses · Scientific methodology

1 Introduction

I am interested in the fact that theory-deduced predictions might change agents' (including, for the most part, human agents) *beliefs*, and thus also agents' *behavior*. Since agents *react* to their beliefs by modifying their behavior to obtain their goals, they might react to a belief inspired by a theory-deduced prediction by modifying

✉ Yotam Harel
yotam.harel@merton.ox.ac.uk

[1] University of Oxford, Oxford, UK

Published online: 04 February 2026

their behavior to obtain their goals, and this may have implications for the theory and its predictive success.

> A classic illustration of this phenomenon is provided by Buck (1963a):
> [A] case, frequently offered as a paradigm in the literature, concerns an agricultural economist's forecast of a future price for wheat. Suppose he forsees an oversupply, and a consequent sharp drop in wheat prices. His prediction comes to the attention of the growers who believe it and decide to switch land to other purposes. So many of them thus switch so much land that the expected oversupply fails to materialize. Perhaps the price even rises a bit. (p. 359)

This phenomenon is quite neglected in the philosophy of science. Except for a limited debate regarding so-called *reflexive predictions* some half a century ago (Buck, 1963a; Romanos, 1973; Vetterling, 1976) following Grünbaum's (1956) comment on Merton (1949), philosophers of science have somehow neglected this phenomenon and its theoretical implications for decades. However, a very recent topical collection entitled "Reactivity in the Human Sciences" in this journal, *European Journal for Philosophy of Science*, marks a possible turning point in this sense, which may, hopefully, ignite interest among philosophers of science in this phenomenon.

Although philosophers of science have dedicated limited attention to this phenomenon, they seem to have diverted the focus from what I regard as its essence, that is, the agents' *reaction* to a belief inspired by some theory-deduced prediction.[1] I consider this reaction essential in the sense that, as I attempt to show in this paper, it might have crucial implications for scientific methodology. For this reason, I suggest that it may be useful to focus on the agents' reaction.

This diversion of focus has led to several misnomers: from *reflexive predictions* (Buck, 1963a; Kopec, 2011; Romanos, 1973; Vetterling, 1976) and *reflexivity* (Beinhocker, 2013; Soros, 2013) to how *performativity* is sometimes used in the context of its implications for predicting (see Mäki, 2013 for a critical formulation of MacKenzie, 2006; Van Basshuysen et al., 2021). These different names can be seen as a mere matter of terminology, but I suggest that they reflect the diversion of focus from the agents' reaction. In what follows, I shall therefore advocate *reactivity* as an adequate name for any theorization of this phenomenon. By doing so, I follow a naming employed mainly by economists (for example, Frey, 2018) but also by contemporary philosophers of science who choose it as some sort of a "general label" over other possible names (Marchionni et al., 2024, p. 2).

In the few instances in which philosophers of science have debated concerning this issue, however, they have not always aimed at the same target. Debaters asserted that their account is justified as it captures what is "important" (Romanos, 1973, p. 104), "significant" (Lowe, 2018), or "matters" (Northcott, 2022, pp. 1, 5), and despite being explicitly stated, these were often different. Indeed, what is important, significant, or matters seems *subjective* relative to what one aims to obtain. Since different

[1] It is important to note at the outset that, somewhat in Lowe's (2021) vein, I hold that the agents' reaction should be inspired by the *content* of a theory-deduced prediction so as to develop a *cognitive attitude* of belief toward it.

debaters attempted to obtain different goals, at times, they were not really debating but rather, as Buck himself puts it regarding his so-called "debate" with Grünbaum, "Grünbaum and I appear to be engaged in a polite exercise of talking past each other" (Buck, 1963b, p. 373). Thus, this paper will do its best to explicitly state its goal and humbly suggest that its focus may prove useful for scientific methodology. Hopefully, this may enable us to avoid these pseudo-debates of the kind characterized here by Buck.

In what follows, then, I shall focus on the phenomenon that theory-deduced predictions might change agents' *beliefs* and thus also agents' *behavior*, and as agents *react* to their beliefs by modifying their behavior to obtain their goals, they might react to a belief inspired by a theory-deduced prediction by modifying their behavior to obtain their goals. This may have implications for the theory and its predictive success. I shall use the general label of *reactivity* in my theorization of this phenomenon since I regard the agents' *reaction* to the belief inspired by some theory-deduced prediction as its essence. I shall dwell on and formulate two distinct kinds of reactivity, which may have significant implications for predicting: self-defeating reactivity and self-fulfilling reactivity. Having set this conceptual framework, I shall examine the implications of self-defeating reactivity for predicting and rethink the so-called exclusivity theses, that is, whether self-defeating and self-fulfilling reactivity is exclusive to the human/social sciences or to states of affairs where human beings/social actors are involved.

This paper's structure is as follows. In Sect. 2, I shall theorize what self-defeating reactivity and self-fulfilling reactivity are. In doing so, I shall consider past attempts to theorize reflexive predictions. First, in SubSect. 2.1, I shall disqualify the popular formulation of reflexive predictions (Kopec, 2011; Romanos, 1973; Vetterling, 1976). Second, in SubSect. 2.2, I shall examine Buck's (1963a) formulation of reflexive predictions and argue that although it has its virtues, it also has some problems. As I consider my account of reactivity adequate relative to what I aim to obtain, in these two subsections, I do *not* claim that these formulations are "wrong" but rather inadequate. In SubSect. 2.3, I shall formulate my account of self-defeating and self-fulfilling reactivity. In Sect. 3, I shall discuss the implications of self-defeating reactivity for predicting. I shall first review Grunberg and Modigliani's (1954) theorem (SubSect. 3.1) and then examine three kinds of self-defeating reactivity: weak (SubSect. 3.2), strong (SubSect. 3.3), and vicious (SubSect. 3.4). In Sect. 4, I shall rethink whether self-defeating and self-fulfilling reactivity is exclusive to the human/social sciences or to states of affairs where human beings/social actors are involved.

2 Self-defeating reactivity and self-fulfilling reactivity

2.1 Reflexive predictions: the popular formulation

In this section, I shall explain what self-defeating reactivity and self-fulfilling reactivity are. However, I shall first show why past formulations of reflexive predictions are inadequate for our purpose. Let us begin by examining the popular formulation of reflexive predictions.

In his theorization of reflexive predictions, Romanos (1973) criticizes Buck's (1963a) formulation of reflexive predictions and advocates an alternative one, which is embraced, for the most part, by Vetterling (1976) and, more recently, Kopec (2011). Due to the popularity of this account in the literature, it will be called the popular formulation. According to Romanos (1973), a prediction is reflexive if and only if its "formulation/dissemination style [is] a causal factor relative to the prediction's coming out true or false" (p. 106). Let us clarify this formulation. The formulation/dissemination style (or F/D-style) is the way in which the prediction is formulated and disseminated. For example, a climate scientist may formulate a forecast according to which it will rain tomorrow in New York in English words and publish it in The New York Times. Thus, the prediction's F/D-style is weather forecast in English/publication in The New York Times. That the F/D-style needs to be a causal factor relative to the prediction's coming out true or false means that the F/D-style needs to "change" the truth value of the prediction at stake. If the F/D-style changes its truth value from false to true, the prediction is self-fulfilling; if the F/D-style changes its truth value from true to false, the prediction is self-frustrating (pp. 106–107).[2]

Vetterling (1976) states that Romanos' formulation is "the most acceptable one" (p. 280), and proposes, on the grounds of his formulation, that a prediction is reflexive if and only if "its dissemination status was a causal factor relative to its falsity (truth)" (Ibid.). In doing so, Vetterling abandons the formulation style and preserves only the dissemination style (or status, in her words). Essentially, except for the exclusion of the formulation style in Vetterling's formulation, both formulations seem identical.

Kopec (2011) adheres to Romanos' account, while, like Vetterling, abandoning the formulation style. According to Kopec's revision of Romanos' formulation, "[a] prediction is reflexive if and only if the mode of disseminating the prediction is a causal factor relative to the prediction's coming out true or false" (p. 1251).[3] Based on these three formulations of reflexive predictions, let us formulate the popular account of reflexive predictions:

> *The Popular Formulation*: a prediction is reflexive if and only if the prediction's dissemination mode is a causal factor relative to the prediction's coming out true or false.

Now, recall that I regard the agents' *reaction* to a belief inspired by some theory-deduced prediction as the essence of what I aim to theorize. Hence, this formulation is inadequate since the prediction's dissemination mode needs *not* necessarily inspire any agents' reaction, and yet, according to this formulation, the prediction may count as reflexive. Although I am not aiming to refute the popular formulation but only show why it is inadequate for our purpose, I think that my argument here,

[2] Romanos explains the causal function of the F/D-style using a theory, T_1, and some alternative theory, T_2 (p. 107). However, this nuance is irrelevant to our purpose.

[3] It is worth noting that Kopec advocates a *weak* formulation of reflexive predictions, according to which the dissemination mode need not change the truth value but only "change the *probability* of the predicted event occurring from what it would be if not disseminated" (italic added, p. 1253). However, this modification of the formulation is irrelevant to our present purpose.

based on counterexamples, might make this formulation's proponents question its correctness.[4]

Let us consider the following example. Assume that some scientist develops a theory that predicts the number of newspaper reports about scientific articles. According to the Theory, T, the predicted number of newspaper reports about scientific articles, P, in some future period t,[5] is Q, so that P_t: Q. Let us also say that Q^* is the actual number of newspaper reports about scientific articles in t. Now, assume that before the scientist published his findings, the theory had been false in the sense that predictions deduced from it prior to its publication had been false. However, the scientist published his theory and deduced a prediction, P_t: Q, in a journal, not intending or even thinking about the possibility that his scientific article would make it to the news. Assume that had the prediction not been published, $Q > Q^*$. But for some reason, many newspaper editors assumed that their readers may find interest in the scientist's prediction, *without* developing any cognitive attitude toward the content of the prediction themselves. Notice that the editors' expectation of readership interest could have been based on the scientist's reputation or on the fact that the prediction concerns a current "hot topic" of public discussion. While these may be good reasons to expect readership interest, they do not even require understanding the content of the prediction, which is much less than forming a cognitive attitude toward it. Eventually, then, many newspapers published newspaper reports about the prediction. Given the number of newspaper reports about it, the scientist's prediction, P_t, turned out true, as $Q = Q^*$. Hence, the prediction was reflexive (specifically, self-fulfilling).

Mutatis mutandis, this example can be modified to illustrate a self-defeating reflexive prediction. Now, assume that the prediction was true had it not been published in that journal so that given P_t: Q, $Q = Q^*$. However, since many newspapers published newspaper reports about it, eventually, the prediction turned out false, as $Q < Q^*$.

Both cases are of reflexive predictions if one is to accept the popular formulation – in both cases, the dissemination mode (namely, the journal publication that led to many newspaper reports about it) was a causal factor relative to the prediction's coming out true or false. However, this example shows why the popular formulation is inadequate for our purpose – in both cases, no agent has reacted to a belief inspired by a theory-deduced prediction. This is because the newspaper editors never developed any *cognitive attitude toward the content of the prediction* – they simply assumed that it may be of interest to their readership. However, the editors did *not* necessarily believe that the prediction was true (nor did they disbelieve it), so the content of the prediction *qua* prediction did not inspire the relevant agents in these cases.

However, the popular formulation's proponents may not accept the requirement that agents must develop (at least) a cognitive attitude toward the prediction's content. Let us then consider another example. Assume that a scientometrics scholar predicts the number of publications in open (that is, not edited and peer-reviewed) scientometrics journals, P (deduced from some theory T), in some future period, t.

[4] It may be worth noting that Lowe (2021) suggests a counterexample to the popular formulation, but Lowe himself seems to doubt his example's ability to convince this formulation's proponents (pp. 82–83).

[5] In this paper, I relate to periods, t_is, in quite a loose sense: They can represent a *moment* or an *interval*, depending on the context.

According to his prediction, there will be Q such publications, so P_t: Q. Let us also say that Q^* is the actual number of publications in open scientometrics journals in t. Now, assume that the scholar published his prediction in an open scientometrics journal. It is known that had the scholar's prediction not been published in that journal, $Q > Q^*$ by one publication only. However, since the scholar published his findings in this open scientometrics journal, his prediction, P_t, eventually turned out true, as $Q = Q^*$. Therefore, the prediction was reflexive (specifically, self-fulfilling).

Mutatis mutandis, this example can be modified to demonstrate a self-defeating reflexive prediction. Now, assume that the scholar's prediction was true had it not been published in that open scientometrics journal so that given P_t: Q, $Q = Q^*$. However, since this scholar had published his work in that open scientometrics journal, eventually, the prediction turned out false, as $Q < Q^*$, once again, by one publication only.

Here as well, both cases are of reflexive predictions if one is to accept the popular formulation – in both cases, the dissemination mode (namely, the journal publication) was a causal factor relative to the prediction's coming out true or false. Once again, this example shows why the popular formulation is inadequate for our purpose, as in both cases, no agent has reacted to a belief inspired by a theory-deduced prediction, not to say developed a cognitive attitude toward the content of a prediction. Furthermore, I suggest that the last counterexample in particular is likely to make the popular formulation's proponents themselves question its correctness as it might seem too broad in the face of these last cases.

2.2 Reflexive predictions: Buck's formulation

Buck's (1963a) formulation of reflexive predictions is, I suggest, more adequate for our purpose than the popular formulation. However, it still leaves some problems to be solved. Let us first present Buck's formulation. According to Buck, a prediction is reflexive if and only if it satisfies the four following conditions:

(1) Its truth-value would have been different had its dissemination status been different,
(2) The dissemination status it actually had was causally necessary for the social actors involved to hold relevant and causally efficacious beliefs,
(3) The prediction was, or if disseminated, would have been believed and acted upon, and finally
(4) Something about the dissemination status or its causal consequences was abnormal, or at the very least unexpected by the predictor, by whoever calls it reflexive, or by those to whose attention its reflexive character is called. (pp. 361–362)

Let us clarify these conditions. Condition 1 entails a counterfactual assessment of what had been the case had the prediction remained private (namely, had not been published) relative to what is the case for a public (namely, published) prediction (pp. 360–361). According to Conditions 2 and 3, the dissemination status (public or private) must be causally necessary for the social actors involved to hold relevant and causally efficacious beliefs, namely, beliefs that made (or would have made)

them react by modifying their behavior. Condition 4 entails that something about the prediction's dissemination status or its causal consequences was abnormal relative to "the usual standard conditions" (p. 361).

Buck's formulation has a lot to offer, and yet, it also has several inadequacies. First, and this reservation applies also to the popular formulation, if the agents' reaction is considered essential, being reflexive must be a *multi-place predicate* rather than a one-place predicate, *Rx*. A prediction is always reflexive only in relation to the relevant agents' reaction, and it is thus senseless to speak of reflexivity as a one-place predicate of predictions. This is also clear from Condition 2 in Buck's formulation, where it is stated that the dissemination status must be causally necessary for the social actors involved to hold relevant and causally efficacious beliefs. Hence, I suggest that reflexivity (or, as I shall call it later on, reactivity) must be a multi-place predicate, where a prediction can only be "reflexive" *relative to the relevant agents' reaction*, rather than a one-place predicate.

Second, and this reservation also applies to the popular formulation as well, the focus on predictions alone misses some epistemic implications that so-called reflexivity has for science. Instead, I suggest that a more adequate formulation should include a distinction between *scientific theories* and *scientific predictions*. Scientific theories are, roughly speaking, *general* propositions about the world. Scientific predictions, on the other hand, are *specific* propositions about the world, typically referring to some *future* period (but always to phenomena that observations on which are at least *unknown* to the predictor),[6] which were *deduced* from some scientific theory (see Lowe, 2021, p. 79 on the deducibility of predictions from theories). To explicate this distinction, let us employ Hempel's notorious proposition, "All ravens are black." This proposition seems a general proposition about the world: It is general by virtue of being universal (about all ravens) and is about the world in being about objects in the world (ravens). One may deduce a prediction from this theoretical proposition. For example, "The next raven John will see will be black." This is a scientific prediction since it is specific (refers to a specific object), refers to some future period (in which John will see a raven), and can be deduced from a scientific theory (from the theory "All ravens are black").

If so, let us distinguish between scientific theories (*T*s) and scientific predictions (*P*s). The relation between them is that every *P* is deduced from some *T*. In the ravens case, for instance, *P*: *Ra*&*Ba* (where *a* is the next raven John will see) is deduced from *T*: $\forall x(Rx \supset Bx)$. This seems to be the case in all the examples discussed above: the weather forecast is deduced from some climate theory; the economic prediction is deduced from some economic theory. Let us then say that for every *P*, there is a *T* for which if *T*, *P* (in other words, *P* is deducible from *T*). This distinction will prove useful when I formulate what self-defeating and self-fulfilling reactivity are, and the lack of this distinction in the formulations reviewed leads to missing some epistemic implications that, I suggest, are worth observing.

[6] Notice that predictions "may be about phenomena that have occurred but observations on which have not yet been made or are not known to the person making the prediction" (Friedman, 1953, p. 9). While predictions referring to a future period may serve extra-scientific practical goals, predictions of the kind Friedman describes are mostly a useful means for testing theories.

Third, Buck's formulation seems contingently restricted to *social actors*, for no apparent reason. In other words, according to this formulation, a prediction is reflexive only when social actors are involved. I consider this a weak point of Buck's formulation since it presupposes *what needs to be proven*: There is a debate as to whether the phenomenon of reactivity is exclusive to states of affairs where human beings/social actors are involved (Grünbaum, 1956, p. 240; 1963; Buck, 1963a, pp. 366–368; 1963b; Romanos, 1973; Vetterling, 1976, pp. 281–282). In formulating this phenomenon as *ipso facto* exclusive to states of affairs that involve social actors, Buck formulates reflexivity in a too-restrictive way that presupposes an answer to this debate.

However, recall that I regard the *agents*' reaction to a belief inspired by a theory-deduced prediction as the essence of what I aim to theorize. Thus, I also seem to restrict the validity of my account – this time to agents. In my defense, I argue that this restriction is justified, and it admits a much wider set of cases. First, it seems plausible that only agents can *react to beliefs*. Second, as opposed to social actors, it seems that at least *de jure*, agents can be *non-human* entities, such as rational aliens, sophisticated robots, or advanced artificial intelligence systems.[7]

2.3 What self-defeating reactivity and self-fulfilling reactivity are

Let us now formulate what self-defeating reactivity and self-fulfilling reactivity are with an eye to Buck's reflexive predictions. Recall that our focus is on the phenomenon that theory-deduced predictions might change agents' beliefs, and thus also agents' behavior. Since agents react to their beliefs by modifying their behavior to obtain their goals, they might react to a belief inspired by a theory-deduced prediction by modifying their behavior to obtain their goals, and this may have implications for the theory and its predictive success.

Notice that I have changed the terminology here from "reflexive predictions" to "Self-Defeating Reactivity" (SDR) and "Self-Fulfilling Reactivity" (SFR). This move has two grounds. First, it reflects what I regard as essential, that is, the agents' *reaction* to a belief inspired by some theory-deduced prediction. Second, in this terminology, we are no longer speaking of a one-place predicate of predictions but rather of a three-place predicate of a *phenomenon*, a relation between a scientific theory, scientific predictions deduced from this theory, and agents' reactions.

However, *reactivity* is a pre-existing concept, and we should thus consider its past formulations. First, this concept has two distinct accounts: (i) reactivity in *data collection* and (ii) reactivity in *the uptake of scientific results* (Marchionni et al., 2024, p. 2). We are here interested in the latter, but first, consider reactivity in data collection. This concept is defined by Espeland and Sauder (2007): "… [T]he methodological concept of reactivity [is] the idea that people change their behavior in reaction to being evaluated, observed, or measured…" (p. 1). Similarly, according to Jiménez-Buedo (2021), reactivity in data collection is "the phenomenon by which subjects in an experiment tend to modify their behavior in virtue of their awareness of being

[7] Notice that Buck (1963a; b) and Grünbaum (1963) discuss whether reflexivity can occur only when human beings are involved. I shall address their discussion in detail in Sect. 4.

under study" (p. 3). Moreover, Runhardt (2021) refers to this account of reactivity as a set of effects affecting measurement: "If a human subject knows they are being measured, this knowledge may affect their attitudes and behaviour to such an extent that it affects the measurement results as well. This broad range of effects is shared under the term 'reactivity'." (p. 1).

Namely, reactivity in data collection is when *x* acts differently than *x* would have acted if *x* had not been studied, evaluated, observed, or measured. On these accounts, reactivity results in a possible methodological flaw in social science experiments (Espeland & Sauder, 2007); it may, under several conditions, threaten the validity of the causal inferences drawn from experimental data (Jiménez-Buedo, 2021); and it may affect measurement results in ways that do or do not undermine the accuracy of the measure, depending on the context (Runhardt, 2021). However, reactivity in data collection is less adequate for our purpose as the latter concerns how *theory-deduced predictions* might change agents' beliefs and behavior, not how being studied, evaluated, observed, or measured might change the relevant individuals' behavior.

Before moving on to our subject matter, just to distinguish it from our account, it is worth mentioning another phenomenon that is often put under the umbrella of reactivity in the uptake of scientific results, that is, so-called *looping effects* à la Hacking. According to Hacking (1995), *human kinds* are kinds employed by the human and social sciences that are exclusive to people in a social setting (pp. 352–353). Hacking explains that, contra natural kinds, such as electrons, these human kinds, such as 'child abuse,' are laden with *values*. For example, while 'electron' has no intrinsic normative load, 'child abuse' has a negative one. Since people are able to react to being labeled, they may react to being classified as a certain human kind that they consider having an intrinsic value, either by endorsing the classification or by resisting it. This gives rise to what Hacking calls 'looping effects:' The process in which individuals react to being classified as a certain human kind, thereby *changing the kind itself*. This creates new knowledge to be acquired about the changing kind, but when it has been acquired – the kind changes again, and so the looping process proceeds (pp. 366–370).

Later, in *The Social Construction of What?* (1999), Hacking introduces the distinction between *interactive kinds* – human kinds, which are exposed to looping effects – that are so named for interacting with what they classify, and *indifferent kinds* – natural kinds, which are unexposed to looping effects (Tsou, 2007, pp. 330–331). The notions of looping effects and interactive kinds appear to have recently gained popularity (see, for example, Khalidi, 2010) and further developments (see, for example, Laimann's (2020) account of capricious kinds).

The phenomenon of looping effects and the human/interactive/capricious kinds driving it concern *how human agents react to being classified as a certain kind*. Indeed, science classifies things, including human beings, into various kinds, from inert gases and mammals to alcoholics and long-term unemployed individuals. However, as Hacking (1995) admits, "[h]uman kinds are formulated in the hope of immediate or future interventions in the lives of individual human beings… The causal understanding (or aspiration to understand) is practical" (p. 351). That is, scientists use human kinds mainly to *predict* the usefulness of future interventions, say, to predict the effectiveness of a treatment for some kind of disease. Thus, in this paper,

I shall begin my inquiry one step ahead – instead of dwelling on how humans' reactions to being classified affect scientific kinds meant to make predictions possible, I shall focus on how agents' reactions to predictions deduced from a scientific theory affect the theory and its predictive success. This is *not* to say that examining looping effects is useless as, among other things, these have an important indirect influence on scientists' ability to predict. Yet, I *do* suggest that my focus on agents' reactions to theory-deduced predictions may prove useful to scientific methodology and is therefore worth considering.

If so, we are interested in reactivity in the uptake of scientific results, focusing on how agents' reactions to predictions deduced from a scientific theory may affect the theory and its predictive success. However, philosophers of science seem to have almost entirely neglected this concept, which is defined and employed, for the most part, by economists (for example, Frey, 2018). Very recently, though, it has been formulated as a general label by Marchionni et al. (2024): "The reactions that science triggers on the people it studies, describes, or theorises about, can affect the science itself and its claims to knowledge. This phenomenon, which we label *reactivity*…" (p. 2). This may serve as a proper point of departure, but it is important to keep in mind that Marchionni et al. formulate reactivity as a general label that includes *both* accounts of reactivity, (i)-(ii). It seems that according to Marchionni et al., reactivity is a phenomenon in which people react to a science that studies, describes, or theorizes about them. This reaction, they hold, can affect this science and its claims to knowledge.

This formulation has many virtues, but some reservations are worth considering at this point. First, notice that this formulation restricts its focus only to sciences that study, describe, or theorize *people*. In other words, this formulation restricts its focus only to the human or social sciences. However, Marchionni et al. stress that "[their] focus here is on the human sciences, in other words the social, psychological, and medical sciences that study people. *This is not to claim that phenomena analogous to reactivity cannot occur in the study of other subject matters*" (italic added, p. 2). This means that Marchionni et al. do *not* claim that reactivity is in fact exclusive to the human or social sciences, and on this point, their formulation merely reflects their focus in this specific paper for a didactic purpose. Therefore, similar to my reservation in SubSect. 2.2, here, we should beware of presupposing *what needs to be proven*: There is a debate as to whether this phenomenon is exclusive to the human or social sciences relative to the natural sciences (Grünbaum, 1956, pp. 239–240; 1963; Buck, 1963a, b, pp. 366–368; Romanos, 1973; Vetterling, 1976, pp. 281–282). We shall, then, avoid restricting our formulation only to "people" science.

Second, for our purpose, the use of "*science*" in the formulation is too broad. What reactions does science trigger? Recall that we are interested in agents' reaction to a belief inspired by some theory-deduced prediction. Thus, let us focus on how science triggers (or, inspires) certain *beliefs*. Science can inspire beliefs in many ways, but here our distinction between scientific *theories* and *predictions* may prove useful.

However, if science can inspire beliefs in many ways, one might wonder why one should focus specifically on theory-deduced predictions. Indeed, this focus has two related grounds. First, the literature on reflexive predictions (Buck, 1963a; Kopec, 2011; Romanos, 1973; Vetterling, 1976) and, to a lesser extent, that on reflexivity

(Beinhocker, 2013; Soros, 2013), seem to focus specifically on predictions. This focus does *not* seem to reflect a mere matter of intellectual taste but rather the belief that predictions are, so to speak, the fruit of science. Since this paper sees itself as a part of this research tradition, which focuses on predictions thanks to their conceived significance for the scientific enterprise, our account of SDR and SFR will focus on theory-deduced predictions to better understand the implications of reactivity for predicting.

However, even if predictions are conceived as such a significant component of science, one might still wonder why one should not formulate a broader account of reactivity that will be able to accommodate *all* the ways in which science inspires beliefs, including predictions. Here, I argue that my strategy will prove useful in analyzing the implications of reactivity for theories and their predictive success: Instead of formulating a catch-all account of reactivity, our formulation of SDR and SFR will be tailor-made for theory-deduced predictions, which will allow us to carefully analyze the implications of reactivity for predicting later on, in Sect. 3.[8]

It is also worth considering Lowe's (2021) account of so-called *self-fulfilling science* to benefit from its merits while carefully distinguishing it from our account of SDR and SFR. According to Lowe, "[a] scientific representation S is self-fulfilling to the degree that S-content-responsive actions contribute to bringing about states of affairs such that a higher degree of conformation between S and S's target system T exists than would have existed in absence of such actions" (p. 91). The main merit of this formulation vis-à-vis our goal is its sensitivity to the fact that so-called self-fulfilling scientific representations (or, SDR and SFR in our case) entail a reaction to the *content* of representations (or, theory-deduced predictions in our case). Restricting this phenomenon to cases where agents react to the content of scientific representations rather than to their other aspects (Lowe, 2021, pp. 82–85), such as their formulation/dissemination style (Romanos, 1973, p. 106), ensures that cases like the newspaper or open scientometrics journal examples from SubSect. 2.1 will remain outside the application range of this concept.

However, Lowe's (2021) account of self-fulfilling science should be carefully distinguished from our account in some important respects. First, in contrast to my account, Lowe's account lacks reference to self-*defeating* phenomena, which are crucial to evaluating the implications of reactivity for scientific theories and their predictive success. That Lowe focuses on self-*fulfilling* phenomena is evident from his book's title, *Self-Fulfilling Science*. More substantially, though, it is clear from the fact that his formulation of scientific self-fulfillment lacks any reference to self-defeating phenomena (p. 91), and that the account does not include a systematic discussion of such phenomena or of how, if at all, the account can be used to accommodate them.[9] On the other hand, my suggested account, as will become clear shortly, applies to both self-fulfilling *and* self-defeating phenomena. It will also become clear, or so I

[8] Just to point out one feature of our formulation that does not necessarily fit all scientific representations, our formulation will rely heavily on the truth value of predictions – propositions referring to specific objects in a specific period – while ascribing, strictly speaking, truth value to other, more general scientific representations, such as models, is often said to be at least problematic (see Lowe, 2021, p. 77).

[9] It is worth noting that, considering the nature of Lowe's account, it is ultimately unclear whether it can simply be "reversed" so as to apply to self-defeating phenomena.

argue, that the current account, tailored to examining, among other things, the implications of self-defeating phenomena for predicting, will prove well-suited for this specific purpose.

Second, it seems that Lowe practically uses a *correspondentialist framework* while resisting employing the *concept of truth*. This is apparent as Lowe states that science is a "representational enterprise… [whose success] depends to some significant extent upon some kind of 'fit' between representational content and states of affairs in the world" (p. 6). On the other hand, Lowe resists employing the concept of truth to characterize this "fit" as he argues that "evaluations of success concerning content-world fit needn't necessarily be cashed out in terms of truth" (p. 78). This results in his use of Longino's (2002) concept of *conformation*, which, on Lowe's (2021) reading, "explicitly encompasses other, more specific, senses such as truth and similarity, among others" (p. 90). In any event, Lowe acknowledges that his account is ill-suited for whoever "recognize[s] an epistemic dimension [of science], but den[ies] that criteria of epistemic success have anything to do with representation" (p. 6). Considering that this characterization seems to correspond to many variants of scientific anti-realists, Lowe's account might be conceived as a nonstarter by them. Instead, my formulation attempts to remain neutral on this matter, and, in this way, it may be of interest to wider audiences of philosophers of science or simply to whoever believes that making successful predictions is (at least) one of the goals of the scientific enterprise. This is because realists and anti-realists alike seem to agree that predictive success is at least *one* of the goals of science, while only instrumentalists seem to believe that it is science's *only* goal (Sober, 1999, pp. 4–5). In this respect, the current account, which does not necessitate a correspondentialist framework (and, as will become clear shortly, avoids relying on questionable counterfactual assessments), forms a genuine alternative that may also appeal to those who position themselves outside the correspondentialist camp.

Let us then formulate SDR and SFR. In doing so, we may consider some points made by Buck (1963a), Marchionni et al. (2024), and Lowe (2021). First, I suggest that from Buck's Conditions 2 and 3 we may accept that these phenomena occur only when relevant agents ("social actors" in his formulation, which has been claimed to be too restrictive) "hold relevant and causally efficacious beliefs" (p. 362). I argue that beliefs here are only causally efficacious when they may lead to a *practical reaction*. Buck seems to share this impression as he states in Condition 3 that the belief needs to be acted upon.

From Marchionni et al. (2024), I suggest embracing the idea that SDR and SFR are *phenomena* rather than merely a one-place predicate of predications (p. 2). This may allow us to formulate SDR and SFR as a relation between scientific theories, scientific predictions deduced from these theories, *and* the reacting agents. From Lowe's (2021) idea of content-responsiveness, we may accept that SDR and SFR occur only when agents react to the *content* of a prediction. Let us then formulate SDR and SFR:

> **SDR (SFR)**: T is self-defeating (self-fulfilling) relative to P_1(s) and P_2(s) of the target v deduced from T and to a set of the agents, $A=\{a_1, a_2, \ldots, a_n\}$, if and only if:

(1) In t_1, P_1(s) was true (false) and in t_2, P_2(s) was false (true),[10]
(2) in t_2, A believe that P_2(s) is true,
(3) obtainment of condition 2 explains A's reaction, R, and
(4) R is causally sufficient for that in t_2, P_2(s) deduced from T was false (true).

Let us clarify this formulation. First, notice that reactivity is now a *phenomenon*, a *relation* of a three-place predicate between a scientific *theory*, a set of scientific *predictions* deduced from this theory, and a set of (the reacting) agents. It is thus no longer a predicate of mere predictions but of theories in relation to scientific predictions deduced from these theories and the relevant agents.[11] This is possible thanks to the distinction between scientific theories and predictions, according to which *T*s are *general* propositions from which one may deduce *many P*s, which are *specific* propositions so that for any *T* there is a *set* of deducible *P*s.

It must be noted, then, that P_1(s) and P_2(s) must be predictions of the same *target*. That is, these predictions must be predictions of the same *variable*, *v*, the theory is used to predict (say, the wheat price in some market or the number of snow leopards in China), but in different *periods* (t_1 and t_2). The distinction between theory and predictions, supplemented by the requirement that the predictions at stake will predict the same target, allows us to use *temporal* assessment instead of questionable *counterfactual* assessment. This seems to be a significant practical advantage for a theory of reactivity that is meant to be of use to scientists. This is because a counterfactual assessment *ipso facto* cannot allow scientists to identify reactivity on empirical grounds: This assessment is speculative as scientists simply *cannot* verify or measure "what had been the case had *x*." Thus, counterfactual assessments of reactivity seem to have limited use to the scientist. In contrast, the inclusion of P_1(s) in our formulation of reactivity may enable scientists to identify cases of SDR or SFR on empirical grounds.[12] When a scientist suspects that a prediction failed or succeeded due to agents' reaction, she can use P_1(s) as a *sanity check* that prevents false-positive identifications of SDR or SFR. For example, after some false P_2s, a scientist might suspect that theory T is self-defeating due to some agents' reaction. Thanks to the possibility of assessing P_1s, even in retrospect, the scientist may undermine this suspicion (when P_1s were also false) or support it (when P_1s were true) on empirical grounds.

[10] Recall that according to Kopec (2011), there is a *weak* sense of reflexivity (p. 1253). It seems, then, that if we embrace Kopec's account, my formulation is of *strong* SDR and SFR, while a revised formulation of condition 1, in terms of probability, may fit as *weak* SDR and SFR. For this paper's purpose, however, the distinction between strong and weak reactivity seems unnecessary, and for convenience, I shall stick to the so-called strong formulation.

[11] It is worth noting that, theoretically, the same theory, T, can be self-defeating in one context (say, for P_1 and P_2) and self-fulfilling in another (say, for P_1' and P_2') relative to the same set of agents, A. This is because SDR and SFR are *not* mutually exclusive for different sets of predictions. However, I struggle to think about actual examples of such cases since there often seems to be clear self-defeating *or* self-fulfilling dynamics in real-world examples.

[12] It may be worth noting that this paper is interested only in cases where the theory was applied *properly*, so it is assumed that the predictions were not erroneously deduced or were no product of mathematical mistakes.

Buck (1963a), on the other hand, asserts that he uses a counterfactual assessment of what had been the case had the prediction remained, in his words, private (namely, unknown to the relevant agents) (pp. 360–361). He does so since he holds that "if the words [in the formulation] are meant in a temporal sense, [there is no] clear meaning in the notion of a prediction, the very same prediction of the very same event, having at one time one truth-value, and at another time, the other" (p. 360). I argue that Buck may be correct as long as there is no distinction between theories and predictions, but when such a distinction is taken into account, Buck's reasoning collapses. Indeed, at least under the traditional commitment to the so-called absoluteness of utterance-truth, a specific prediction, P_t, is always true, false, or neither true nor false (see MacFarlane, 2003, pp. 327–328 for elaboration).[13] However, when it comes to theories, *T*s, which are general propositions, there is actually a clear sense in speaking in temporal terms. Assume that P_1 and P_2 are deducible from T, when P_1 refers to t_1 and P_2 to t_2. P_1 is "It will rain on Sunday" and P_2 is "It will rain on Monday." If it did rain on Sunday but *not* on Monday, P_1 is true and P_2 is false. In this very sense, T was used to derive a correct prediction of rain on Sunday (t_1) but also an incorrect prediction of rain on Monday (t_2). Hence, there is really a sense in speaking in temporal terms as the same theory, T, may be used to derive a correct prediction, P_1, in t_1 but also an incorrect prediction, P_2, in t_2.

To sum up, while the existing alternative formulations of reactivity (or so-called reflexive predictions) are committed to counterfactual assessments, my formulation need not rely on such questionable counterfactual assessments that, *a priori*, can never be verified, thereby avoiding this issue altogether. According to my formulation, we can always verify whether or not condition 1 obtains – we can always test (assuming that we have the relevant data), whether in real-time or in retrospect, whether any P_t is, or was, true or false.

Another advantage in distinguishing between *T*s and *P*s is that agents can, at least in some cases, deduce a P from a T. For example, a scientist may publish the theory "All Sundays are rainy," and if the theory is believed, agents may deduce the prediction "Next Sunday will be rainy." In this sense, it can sometimes be sufficient that a T will be public (and credited) but no *P*s deducible from T will be explicitly published for reactivity to occur.

Notice also that my formulation does not rely on any notion of *dissemination mode* (the popular formulation) or *dissemination status* (Buck's formulation). Instead, in condition 2, I state that P_2(s) is believed to be true by the relevant agents, A. Here, what matters is whether or not these agents *believe* that the predictions deduced from some theory are true.[14]

[13] Notice that this paper need not take sides in the future contingents debate since the assessment of the truth value of predictions relevant to our position takes place only *after* the predicted event occurs (or not). That is, our position is not interested in assigning truth value to predictions before the period referred to by them.

[14] It may be worth noting that Lowe (2021) holds that it is not necessary for the prediction (or, in his account, the scientific representation) to be believed by the agents and supports this position with two peculiar examples, which he himself admits do "not represent what we think of as a paradigmatic action leading to self-fulfillment" (p. 86).

The improvement in formulating things this way is, I think, clear – as I showed in SubSect. 2.1, the notion of *dissemination mode* is inadequate in being too *broad*; Buck's notion of *dissemination status* is inadequate in missing a host of cases and, correspondingly, being too *narrow*. What Buck misses is that agents may *change their minds* regarding a theory's predictions' truth value – a theory can be public, and yet discredited by the relevant agents; predictions can be public, and yet disbelieved by the relevant agents. In these cases, the question is not whether the theory or its predictions are private or public, as Buck puts it, but rather credited (namely, believed to be correct) or discredited (namely, believed to be incorrect). It is clear that also for Buck, if a prediction is public but discredited by all, it cannot be reflexive as it will not be, in his words, acted upon. Thus, what seems to be essential is that the relevant agents come to believe at some point that the theory's predictions are true, whether this change in the agents' attitude results from a publication *or* from a mere change of mind as to the theory's predictions correctness.

As to A, the set of relevant agents, it is a finite set of agents that may range from one agent to many. This set is relevant in the sense that its reaction, R, is causally sufficient for condition 1 to obtain. A should then include enough agents so that their R will be causally effective (as mentioned, in some cases, one agent may suffice).

Let us now exemplify how my formulation applies to a classic example. Recall Buck's (1963a) example of predicting wheat price. Assume that an agricultural economist develops a theory, T, for predicting wheat price in some market. When the economist tests T in t_1, before publishing it, he finds out that its predictions are correct – that every P_1 turns out true. Then, the economist publishes T so that in t_2, a set of wheat growers come to believe it. Now, they deduce that according to T, the predicted wheat price for some period is P_2: p, which reflects an oversupply and, correspondingly, a drop in wheat price. The wheat growers react to the belief that P_2: p and, consequently, produce less wheat. Eventually, since the wheat growers produce less wheat in t_2, there is a rise in wheat price in t_2 so that if p^* is the actual wheat price in t_2, $p<p^*$ and P_2 turns out false.

This is, of course, a case of SDR as P_1s were true and P_2(s) was false. A, the set of relevant agents here, are the wheat growers that believe that T is correct and react to this belief. Their reaction, R, is reducing wheat production, which is causally sufficient for P_2 to be false – this reaction leads to a shortage in wheat and, correspondingly, to a rise in wheat price so that $p<p^*$ and T's prediction, P_2, turns out false.

3 Predicting under reactivity

3.1 Grunberg and Modigliani's theorem

In this section, I shall examine which problems reactivity poses to predicting and, specifically, whether predicting under reactivity is possible. In other words, I aim to figure out whether reactivity poses any *special* problems to predicting. Thus, to figure out whether reactivity poses any special problems to predicting, let us assume that a perfect non-reactive prediction is possible to isolate the impact of reactivity (like Grunberg & Modigliani, 1954, p. 465).

This question has so far received primarily economists' attention. Grunberg and Modigliani (1954) famously proved that *under several assumptions*, predicting under reactivity is always possible. Endorsing their proof, Simon (1954) showed how their general proof applies to predicting election results. Buck (1963a, pp. 364–365) followed Grunberg and Modigliani's Theorem (GM-T) and argued that so-called reflexivity poses no special problems for predicting. Since GM-T is logically impeccable, the only way to challenge it is by challenging its *assumptions*. Indeed, many have successfully shown that some of GM-T's assumptions are rather unrealistic in many cases (Aubert, 1982; Henshel, 1995; Øfsti & Østerberg, 1982). Henshel even goes further and states that the GM-T might be a logically impeccable theorem about an empty set (pp. 517–518). Besides these explicit attempts to question the GM-T's assumptions, I shall suggest that we may read Frey (2018) as pointing to new (and, I suggest, promising) grounds for questioning the GM-T's validity. So, let us employ our formulation of SDR and SFR, with an eye to the GM-T, to examine whether predicting under reactivity is possible.

First, it is clear that SFR poses *no* special problems to predicting since in SFR, the prediction(s) must be *false* in the first place although we must assume that a perfect non-reactive prediction is possible. In other words, reactivity can be said to pose special problems to predicting only when in t_1, P_1(s) is true. However, in SFR, P_1(s) are *ipso facto* false. Besides, recall that in this section, we are interested in the *problems* reactivity may pose to predicting, namely, in how reactivity may make predicting more difficult and complex, or even impossible. However, as SFR is self-*fulfilling*, it can only help us predict[15] rather than posing problems to predicting. Hence, reactivity can pose special problems to predicting only as a result of SDR.

So, let us now explicate GM-T to see whether its assumptions hold while establishing the analysis on our formulation of SDR. The GM-T assumes that a public (or, in our terms, a reacted-upon) prediction of a variable is possible only if there is an *equilibrium* in which the predicted value of the variable, P_2, is equal to the actual value of this variable in this period, given the agents' reaction, P_2* (p. 472). So, we have a first equation: (1) $P_2=P_2$*. Now, assume that we know to define the reaction function of the relevant agents – the actual, "reacted-upon" value of the variable as a function of any prediction of this value (p. 471). Let us call this reaction function $R(P_2)$. We then have a second equation: (2) $R(P_2)=P_2$*. Now, the question is whether there is a P_2 for which, given $R(P_2)$, $P_2=P_2$*. Here, a graphic explication of the question may help (see Fig. 1).

Figure 1 shows that such an equilibrium exists if and only if $R(P_2)$ intersects with $P_2=P_2$*, which is, graphically, the 45° line through the origin. Now, according to Brouwer's Fixed Point Theorem, such an equilibrium exists if the predicted variable possesses a lower bound k and an upper bound K, where k and K are real and finite, and the function $R(P_2)$ is continuous over the interval $[k, K]$ (Grunberg & Modigliani, 1954, p. 472). Grunberg and Modigliani hold that since "[t]hese conditions were

[15] I use 'predict' in the simplest sense of deducing a proposition (that is, a prediction) that says what the value of some variable v, unknown to the predictor when making the prediction, will be (or was) in some t. In this sense, a successful prediction only needs to prove correct in t. I stress this just to put aside questions that often preoccupy philosophers and concern the correspondence between the theory (the prediction was deduced from) and reality.

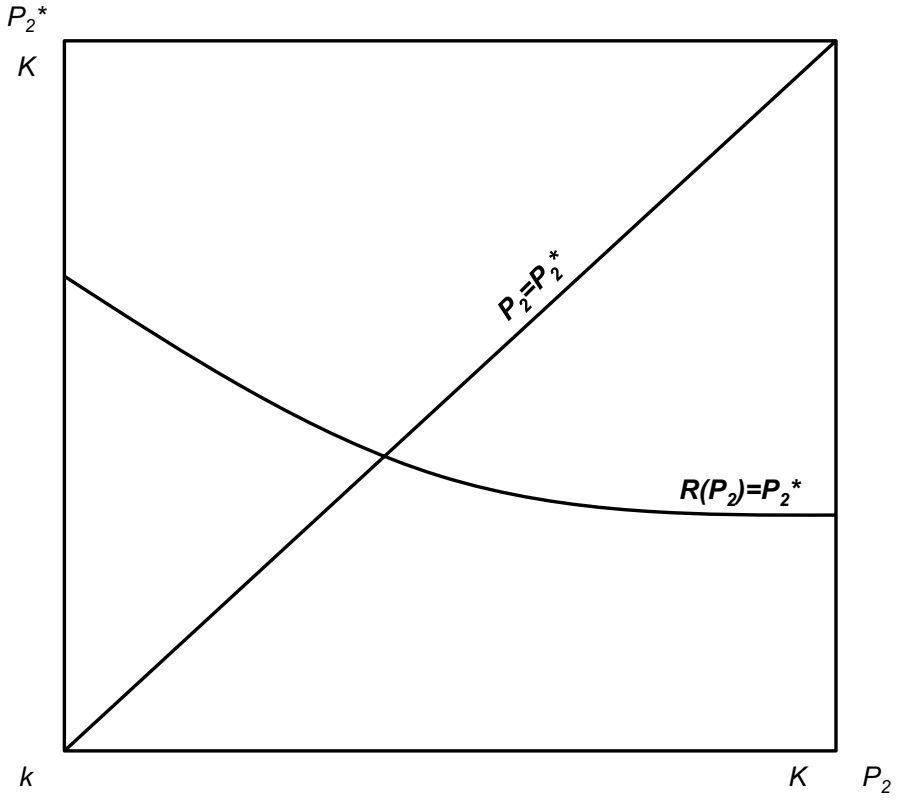

Fig. 1 Presents a case in which a true reacted-upon prediction is possible, where the horizontal axis describes the range [k, K] of the predicted value of the variable, P_2, the vertical axis describes the range [k, K] of the actual, reacted-upon value of the predicted variable, P_2*, the 45° line through the origin represents the equation $P_2=P_2$*, and the curve represents the equation $R(P_2)=P_2$*

found to be normally fulfilled in the world about which predictive statements are to be made" (p. 478), a true reacted-upon public prediction is *normally* possible (p. 469).[16]

If so, SDR has been proved to pose *no* special problems to predicting only under the following assumptions, assumed in GM-T:

(1) *Boundedness Assumption*: The predicted variable possesses a lower bound k and an upper bound K, where k and K are real and finite.
(2) *Continuity Assumption*: $R(P_2)$ is continuous over the interval [k, K].
(3) *Definability Assumption*: $R(P_2)$ is definable over the interval [k, K].[17]

[16] As Grunberg and Modigliani show, their theorem can be also expanded to predicting many variables (pp. 473–474), but this is irrelevant for our purpose here.

[17] Assumption (3) is necessary for assumption (2) so that if assumption (2) obtains, assumption (3) obviously obtains as well. I distinguish between these assumptions to clarify the distinctions between the different kinds of reactivity.

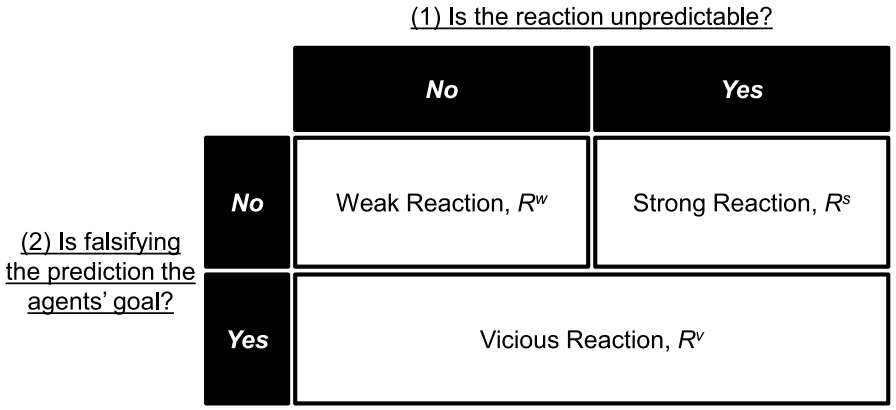

Fig. 2 Presents the typology of reaction kinds according to two questions: (1) is the reaction unpredictable, and (2) is falsifying the prediction the agents' goal, and the corresponding typology to weak, strong, and vicious reaction

In the following subsections, I shall attempt to show that assumptions (2)-(3) do *not* hold in many cases. To distinguish between these cases, I shall distinguish between different kinds of reactions, *R*s: *weak*, *strong*, and *vicious* (see Fig. 2). In SubSect. 3.2, I shall show that weak SDR might make a reacted-upon prediction impossible in many cases, due to a violation of the continuity assumption. In SubSect. 3.3, I shall show that strong SDR makes a reacted-upon prediction impossible, due to a violation of the definability assumption. I shall argue that what leads to the problem in this case is not exclusive to cases of SDR, and yet reactivity can trigger the problem. In SubSect. 3.4, I shall show that vicious SDR makes a reacted-upon prediction impossible, due to a (deliberate) violation of the continuity or the definability assumption.

3.2 Weak SDR

Now, let us consider cases of weak SDR. As we distinguish between kinds of SDR by virtue of different kinds of *reactions*, let us first explain what a weak reaction, R^w, is. So, a reaction is weak if and only if (1) the reaction is *predictable*, and (2) falsifying the prediction at stake is *not* the agents' goal. Such reactions are implicitly assumed, for the very most part, in the literature regarding reactivity (or so-called reflexivity).

Condition 1 above seems plausible to assume here since we assume, for the sake of this discussion, that a perfect non-reactive prediction is possible. This is because

> … once [non-reactive] prediction is assumed to be possible, the agents' reaction to a public prediction must also be regarded as knowable. For the assumption that private prediction is possible implies that it is possible to ascertain (*a*) how the agents' expectations are formed and change as a result of given information and (*b*) how the agents act in response to given expectations. (Grunberg & Modigliani, 1954, p. 466)

Condition 2 above also seems plausible in most cases since agents do not seem to attempt to falsify scientific predictions regularly. What shall then be argued here is that weak SDR might make a reacted-upon prediction impossible in many cases due to a violation of the *continuity assumption*. To do so, I shall rely on Aubert (1982), Øfsti and Østerberg (1982), and Henshel (1995) and show that for many phenomena, the GM-T does *not* hold and a true reacted-upon prediction might be impossible.

First, consider the case of predicting *election results*. This is, of course, one of the most central goals of political science. However, the continuity assumption does not hold here. First, elections take place within finite populations of n members, where the candidate who gets the highest percentage of votes wins. Now, if the number of votes to the candidate i is v_i, the winning candidate will have the highest v_i/n. However, the problem is as follows: The variable to be predicted, v_i/n, is not continuous *in itself*, and hence $R(P_2)$ cannot obviously be continuous over $[k, K]$, which is here [0, 1] (Aubert, 1982). Now, an actual example may illustrate the problem. Assume that there is a mayoral election in a town with 10 eligible voters (when all voters indeed vote) and only 2 candidates. Now, assume that a political scientist attempts to deduce from some T a P of the percentage of votes candidate 1 will have. Recall that this variable is bounded between [0, 1], or [0%, 100%], which may be more illustrative. Now, $R(P_2)$ is continuous over the interval [0%, 100%] only if it is defined at *every* point on it. But the problem is that the interval itself is not continuous so that $R(P_2)$ *cannot* be continuous over it. In fact, the predicted variable, v_1, can have only 11 values: 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%. v_1 cannot be equal, for example, to 25% or 62.4%. Hence, as the predicted variable is not continuous over the relevant interval, so must be $R(P_2)$, and the continuity assumption does not hold.

Another problem in the case of predicting election results is as follows: Predictions are regularly presented to and perceived by the voting agents in *rounded-off numbers*. For instance, if some candidate is predicted to have 42.8752% of the votes, the prediction will likely be presented to the voters as 42.9% or even 43%. Even if it is presented to the voters in its precise form of 42.8752%, most voters will perceive it as the same as 42.9% or even 43%. Since Ps of a percentage of votes are regularly presented and perceived as round numbers (or with very few figures after the decimal separator), $R(P_2)$ is obviously not continuous over the interval [0%,100%] (Øfsti & Østerberg, 1982).

So, it appears that the continuity assumption does not hold when the predicted variable is the percentage of votes in an election. Hence, SDR might make a prediction of election results impossible in many cases, and in these cases, of course, GM-T does not hold.

Second, consider the case of predicting *prices*. There can barely be a more fundamental concept in neo-classic economics than the concept of price. Here as well, the continuity assumption does not seem to hold in reality. This is because consumers are regularly focused on the left-most digits, which serve as benchmarks for decision-making. Hence, there is a discontinuity around the change of a *left-most digit*. As Henshel (1995) puts it,

> [f]or instance, a retailer prices a commodity at $39.99 instead of $40… If this is as nearly universal as I believe, what is the reason for such a practice?... The point is that, if the retailers are correct, there is a *discontinuity* at the value $40 in the relationship between price and quantity sold. Prices of $40, $41, $42, and so on – for this particular product – do not represent simple extensions of the demand curve represented by... $37, $38, $39, $39.50, $39.90.... A significant proportion of consumers resists purchasing the product above $40, which forms a benchmark value. Were we to graph the relationship, a discontinuity would appear at $40 – and probably another at $50. (p. 510)

If so, it seems that $R(P_2)$ cannot be continuous when the predicted variable is a price as the consumers' demand for a product is discontinuous around changes in the left-most digits. Thus, it appears that weak SDR might make a reacted-upon prediction impossible in many cases due to a violation of the continuity assumption (see a graphic illustration of such a case in Fig. 3). I reviewed here only two phenomena

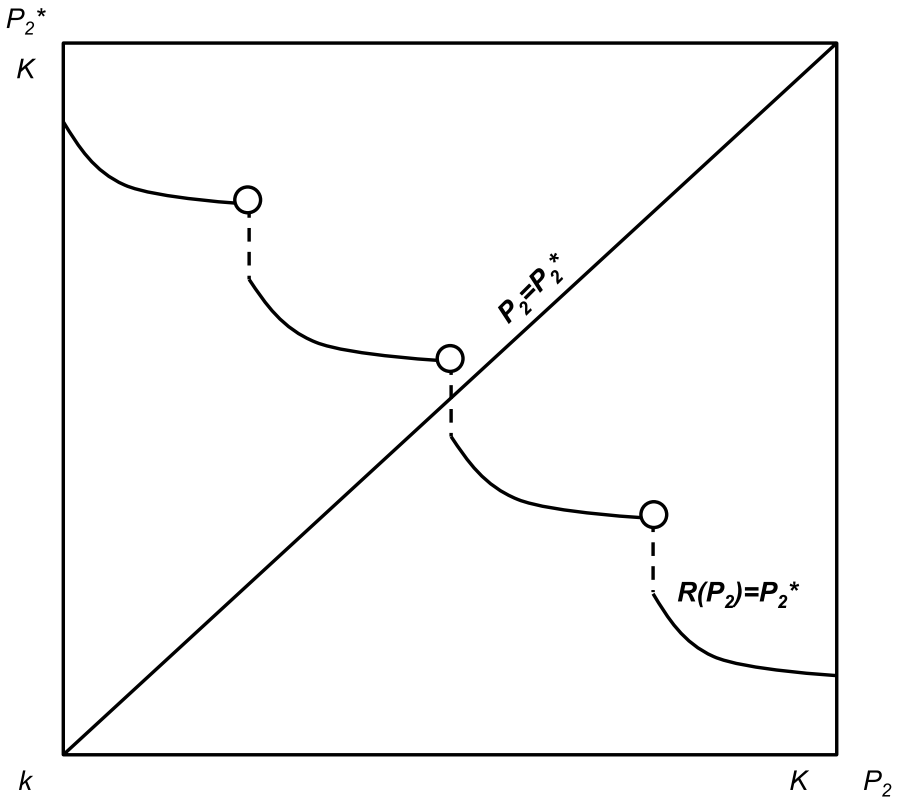

Fig. 3 Presents a case in which a true reacted-upon prediction is impossible due to a violation of the continuity assumption, where the horizontal axis describes the range [k, K] of the predicted value of the variable, P_2, the vertical axis describes the range [k, K] of the actual, reacted-upon value of the predicted variable, P_2*, the 45° line through the origin represents the equation $P_2=P_2$*, and the discontinuous curve represents the equation $R(P_2)=P_2$*

in which it seems that this assumption is violated due to space limits, but considering the centrality of these phenomena to their disciplines (election results to political science and prices to economics), these examples suffice at this point. Then, weak reactions, R^ws, *might* pose a problem for the possibility of predicting. However, it is worth noting that they do not necessarily lead to such problems in *all* cases. My argument here advocates two conclusions. First, the continuity assumption does *not* hold in many cases so that the GM-T does not hold in many cases, and the question as to the existence of an equilibrium where $P_2=P_2$* given a weak reaction remains open. Second, since the continuity assumption does not hold in many cases, and when it does not hold, a true prediction might be impossible, correspondingly, weak SDR *might* make a reacted-upon prediction impossible in many cases. Hence, in these many cases, a case-specific examination is needed to figure out whether a reacted-upon prediction is even possible, and the answer might turn out to be negative.

3.3 Strong SDR

Strong SDR is *strong* in two senses. First, it is strong in the sense that in contrast to weak SDR, when it occurs, a prediction is *always* impossible, not only *might* be so. Second, it is strong in the sense that it entails a strong reaction, R^s. I call a reaction strong if and only if (1) the reaction is *unpredictable*, and (2) falsifying the prediction at stake is *not* the agents' goal. If so, a true reacted-upon prediction is always impossible in cases of strong SDR as the definability assumption never holds and $R(P_2)$ is undefinable so that scientists cannot predict the agents' reaction to predictions.

But why should any agents' reaction be *unpredictable*? Here, I suggest that the answer is *creativity*: Agents (or at least some of them) react to beliefs in creative ways. Creative reactions are unpredictable since no data preceding the moment of creative reaction could have led to the conclusion that the creative reaction will take place.

So, R^ss make it impossible to predict in cases of SDR. As Frey (2018) puts it,

> [i]t seems impossible to foreseen all kinds of reactivity. Human beings often exhibit immense creativity to react to outside interventions. Innovative reactions are by definition not predictable – not even by a "Master Algorithm" – because they would otherwise have already been used. Whether a learning machine can foresee such creative reactions boils down to how innovative human beings are considered to be. In my view they are creative and able to find completely new ways to react. If this were the case, not even a very extensive game theoretic model would suffice. (p. 148)

To this extent, I agree with Frey. I shall now provide an example of such a case of strong SDR. Assume that a physicist uses T, a predictively successful theory in some t_1, to predict the motion of an asteroid. The physicist deduces P_2 of T, when P_2 is as follows: "The asteroid will hit Earth in another 100 days." This prediction is based on the physical theory *and* on the *correct* assessment that no existing technology at the moment of predicting P_2 can prevent its collision. The worried physicist immedi-

ately publishes P_2, which is widely believed by Earth residents. However, a brilliant engineer is not willing to give up, and he does his best to invent a technology that can somehow prevent the predicted collision. He then strongly reacts, R^s, to P_2 and invents a technology that can prevent the disaster. Thanks to his invention, the asteroid does *not* hit Earth ultimately.

A twofold note is worth considering at this point. First, this case is a clear example of SDR as conditions 1–4 obtain. Second, this example shows that reactivity is *not* exclusive to human/social sciences: The prediction here is deduced from a natural science theory. However, this will be thoroughly discussed in Sect. 4.

Now, an important remark must be made. If we accept that at least some agents can *react* creatively, it seems only plausible that they can also *act* creatively. Namely, agents can act in creative ways also when they are not reacting to a specific prediction. Thus, it is impossible to predict creative acts also without reactivity so that the problem of predicting given creativity is not actually exclusive to reactivity.[18] Indeed, this problem is not exclusive to cases of reactivity, but my example may be taken to show that predictions may *trigger* creative acts that otherwise, seem very unlikely. It seems very unlikely that the engineer would have invented his novel asteroid-collision-preventing technology unless an asteroid has been predicted (and believed) to hit Earth. In this sense, predictions in cases of strong SDR may trigger creative reactions that would not have taken place (in the form of an act) had the prediction not been believed.

If so, in strong SDR, the definability assumption is violated and a true reacted-upon prediction is never possible. The problem is that since a strong reaction is unpredictable, scientists cannot predict in which cases strong SDR will emerge, in contrast to weak SDR that can be predicted (even when it is predicted that given weak SDR in some case, a reacted-upon prediction is impossible). Indeed, I stress again that the predictive difficulties that strong SDR raises are not exclusive to reactions in the sense that if we assume that creativity is possible, a non-reactive prediction might also be impossible in some cases, as agents can act creatively also when they are not reacting to a certain prediction. However, I do argue that reactivity may trigger creativity so that it may increase the probability of an unpredictable act.

3.4 Vicious SDR

Vicious SDR is vicious as its reaction is vicious. A reaction is vicious, R^v, if and only if falsifying the prediction at stake *is* the agents' goal. I shall show that whether the reaction is predictable or not is irrelevant here since the agents always react to the prediction only *after* the scientist has made it. In any event, when the reaction is predictable, the continuity assumption is deliberately violated – the agents, so to speak, "jump" over the 45° line so that an intersection cannot be found. When the reaction is unpredictable, the definability assumption is violated.

[18] It may be worth noting that this is not to say anything about the *prevalence* of creative acts – in fact, it seems that most of our acts are *not* genuinely creative (say, our behavior is not typically very creative when we drive to work, choose products at the grocery store, and the like). Thus, we can (and indeed do) accurately predict human behavior in many cases where agents do not regularly act creatively.

Grunberg and Modigliani (1954) think about the possibility of what I call a vicious reaction but also state that it "does not seem to be empirically important" (p. 475). They hold that where agents aim to falsify a (public) prediction, a prediction is impossible, since "[t]he situation here becomes that of a game of matching coins where, however, one player must show his coin before his opponent decides what move to make: correct public prediction is here impossible" (Ibid.).

In what follows, I shall suggest a more formal argument than the coins matching game and prove that in cases of vicious SDR, a true reacted-upon prediction is never possible. Afterward, I shall argue that Grunberg and Modigliani's claim that vicious SDR is empirically unimportant might be somehow hasty.

Let us now model vicious SDR as a game. Assume that: (1) the goal of the scientist is to predict the value of the variable v, (2) the goal of the agent is to falsify the scientist's prediction, (3) the agent can determine the value of v, (4) and the game is a two-stage game, where the agent chooses a strategy after the scientist chooses a prediction (and publishes it). Let us say that the scientist is player 1 and the agent is player 2 so that if the scientist predicts v, $U_1=1$ and $U_2=0$ (1, 0); if the scientist fails to predict v, $U_1=0$ and $U_2=1$ (0, 1). So, in the first stage, the scientist chooses some P, a prediction of v. In the second stage, the agent chooses the value of v. For any P, the agent has two options: (1) choose $v=P$ and the payoff (1, 0); (2) choose any $v\neq P$ and the payoff (0, 1). Since the agent, player 2, prefers (0, 1) over (1, 0) for any P chosen by player 1, in every subgame perfect equilibrium $v\neq P$ – namely – a prediction is *not* possible (see Fig. 4).

But if it is merely a game theoretic quibble, Grunberg and Modigliani are correct in saying that vicious SDR is empirically unimportant. I suggest that although it seems less frequent than weak SDR or strong SDR, vicious SDR may occur in some cases of *strategic skepticism*. Vicious SDR is likely to occur when agents can benefit from undermining the *authority* of a predictor. When agents can benefit from it, they might be willing to bear some costs in the *short* term to undermine the predictor's authority in the *long* term. Let me exemplify. Recall our wheat price example from SubSect. 2.3. However, assume now that the wheat price in the country is controlled by the chief economist of the Ministry of Agriculture. The economist uses some T to predict the consumers' demand and the supply produced for any given price, p. During some t_1, the economist's P_1s were accurate: There was no oversupply or shortage in wheat. However, the wheat growers were dissatisfied with their profit and claimed that the controlled price was too low. But since the economist's predictions turned out true in t_1, their claim was not convincing.

In t_2, they decided to act – they decided to undermine the economist's authority by falsifying his prediction. The economist predicted some P_2: p, but the wheat growers produced *less* wheat than the economist had predicted. The decision to produce less wheat was not optimal, profit-wise, *in this period*, and yet – they were successful in falsifying the economist's prediction: Eventually, there was a shortage of wheat. Now, their claim that the controlled price was too low may sound convincing – if, given some controlled p, there is a shortage of wheat – the controlled price seems too low. Undermining the economist's authority may lead the minister to intervene and, for example, raise the controlled price.

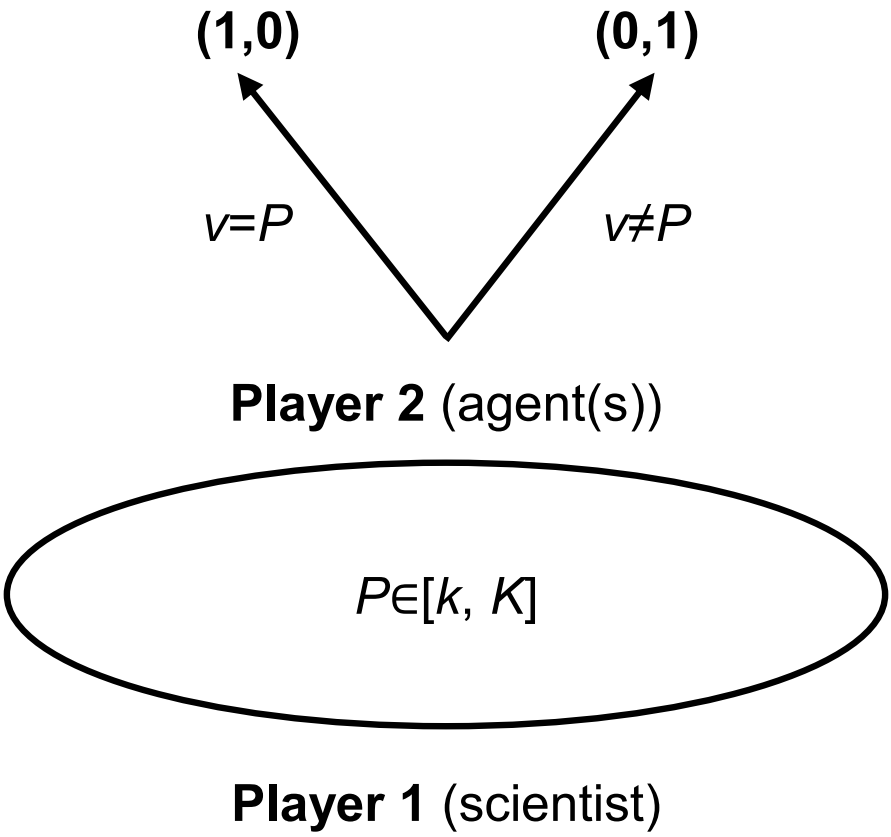

Fig. 4 Presents the game-theoretic model of the scientist and the vicious agent(s), where at the first stage, the scientist chooses any prediction, P, over $[k, K]$, and at the second stage, the agent(s) choose the actual value of the predicted variable, v, where they can choose any value over $[k, K]$, whether $v=P$ or $v\neq P$

This example is not only a theoretic scenario, but what might be a prevalent real-world practice of firms that operate in price-controlled markets. It is well-known that price controls are often followed by *shortages* of the price-controlled products, resulting in the effective price being *raised* (Mitchell & Azevedo, 1975, p. 571). Just recently, for example, Hungary has set price controls on several food products (Spike, 2025), although in past years, such measures in Hungary have led to shortages (Csonka, 2024), pushing the effective price upwards. Now, in order to explain this issue, let us begin by explaining what price controls are meant to do. In markets with imperfect competition, such as monopolistic or oligopolistic markets, prices are often higher than the optimal, efficient level. In an attempt to bring about a more efficient allocation, governments sometimes set price caps to limit the price and bring it closer to the optimal level. However, as mentioned here, this often leads to shortages, which reasonably lead to a higher price later, either by raising or abolishing the price cap. Economists have suggested several explanations for the phenomenon that price

controls are often followed by shortages (see, for example, Mitchell & Azevedo, 1975).

Now, our account of vicious SDR can provide another explanation for this phenomenon: Shortages following price caps can be *deliberate* – firms may intentionally create shortages in a certain period to raise the price in the next periods. This strategy, involving strategic skepticism, enables firms in markets with imperfect competition to effectively evade the intended effect of price controls and maximize profits using the mechanism indicated above. However, it is worth noting that while this explanation for the discussed phenomenon seems reasonable in many cases, it does *not* rule out other explanations for this phenomenon. Moreover, it seems that verifying cases in which vicious SDR has led to shortages is difficult since firms would attempt to *conceal* such manipulations and keep them under wraps.

If so, in vicious SDR, falsifying the prediction at stake is the agents' goal, and when this is the case, a true reacted-upon prediction is always impossible. While at first glance, it might seem that vicious SDR is merely a game-theoretic quibble that never occurs in reality, I have provided an example in which, I think, the occurrence of vicious SDR seems rather plausible. Indeed, it seems that vicious SDR may explain cases in which price controls are followed by shortages. I suggest that although this kind of reactivity might seem relatively rare, it may occur in cases of strategic skepticism and is thus not empirically unimportant, as Grunberg and Modigliani argue. On the contrary, this kind of SDR deliberately endangers the epistemic authority of scientists and scientific knowledge and should therefore be observed carefully.

Now, I shall add a remark regarding these kinds of SDR and Northcott's (2022) account of *fragility*. According to Northcott, "a relation [is] *fragile* if, in the salient circumstances, it is not predictable" (p. 2). As to the relation between fragility and so-called reflexivity (reactivity in our terms), Northcott argues that "reflexivity can be a useful indicator of fragility" (p. 11). I think that my formulation of reactivity and of the three different kinds of SDR may *explain* why reactivity is an indicator of fragility, so defined – as we have just seen, SDR may render a certain phenomenon *unpredictable*. Hence, SDR may cause unpredictability. Thus, if a relation is fragile when it is unpredictable, it is clear why SDR can make phenomena fragile. In this way, my formulation of reactivity may explain why and how some relations are fragile.

4 The exclusivity theses

In this section, I shall examine what I call the *exclusivity theses*. I say theses in the plural since this discussion has so far suffered from a lack of clarity. A distinction between two distinct yet interrelated exclusivity theses, which is in line with Lowe's (2021) related convincing argument (pp. 97–103), may aid in clarifying this discussion and show, as I argue, that with the aid of the distinct, clarified theses, any perplexity vanishes. It is worth noting, however, that while Lowe's strategy seems similar to mine, my inquiry examines the exclusivity theses with respect to an entirely different thing: While Lowe examines these theses with respect to his account of self-fulfilling science, my inquiry focuses on my account of SDR and SFR. That is, while both of us attempt to figure out whether '*x* is exclusive to the human/social sciences *or* states of

affairs where human beings/social actors are involved,' since our accounts of x differ, our inquiries naturally aim at different targets.

One question is whether or not reactivity is exclusive to the *human/social sciences*. As Marchionni et al. (2024) put it clearly, "reactivity has often been taken to mark a significant ontological and epistemological difference between the human and the natural sciences" (p. 2). The thesis at stake here, then, may be formulated as follows:

> *Exclusivity Thesis 1* (ET1): The phenomenon of reactivity is exclusive to the human/social sciences.

Another question is whether or not reactivity is exclusive to the so-called "social realm" (Lowe, 2018, p. 344).[19] Grünbaum (1956) asks whether reflexivity is exclusive to "the realm of human affairs" (p. 240). Marchionni et al. (2024) sketch the debate between Grünbaum (1956, 1963) and Buck (1963a, b) in terms of whether or not there is a "difference between the social and the natural worlds" (p. 8). However, the talk of the social/human realm/world sounds to me admittedly metaphorical and rather unclear. Buck (1963a) formulates the question in a clearer manner and asks "whether reflexive predictions occur only when people are involved" (p. 366). If so, let the second thesis be:

> *Exclusivity Thesis 2* (ET2): The phenomenon of reactivity is exclusive to states of affairs where human beings/social actors are involved.

Now, when the theses are distinct and clearly defined, I suggest that questions regarding their correctness dissolve if we follow our formulation of reactivity. First, let us examine ET1. The human/social sciences are psychology, sociology, economics, and the like. The natural sciences are physics, chemistry, geology, and the like. Now, one example of reactivity in one of the natural sciences suffices to refute ET1. In other words, if we find at least one case in which some theory of the natural sciences turns out reactive, ET1 proves false. It seems, then, that we already have an answer – recall our asteroid example from SubSect. 3.3. In this case, T, from which P_2 is deduced, is taken from the *natural sciences* – from *physics*. The prediction is deduced from a physical theory that theorizes the motion of bodies. So, if the asteroid example is adequate, it is clear that ET1 is false as we have an example of reactivity in the natural sciences.

But it is surprisingly easy to think about more counterexamples to ET1. I shall review here only one more example due to space limits. Assume that some zoology theory predicts the extinction of animal species. In t_1, the T's P_1s were perfectly accurate. In t_2, a scientist predicts P_2, deduced from T, according to which some species will go extinct in several years. This prediction is believed, and for some reason, relevant agents, A, react to obtain one common goal, that is, to save the species from extinction. Their conservation efforts, R, bear fruit, and as a result, the species is saved so that P_2 is false. This description seems to fit many real-world examples in which a species was predicted to go extinct, but due to conservation efforts in reac-

[19] Notice that Lowe does *not* discuss this question there.

tion to this prediction, the species was eventually saved. For example, some decades ago, many scientists believed that the snow leopard (*Panthera uncia*) was likely to soon go extinct (see, for example, Jackson, 1979, p. 194; Schaller et al., 1988; Munson & Worley, 1991, p. 37). However, governments and other organizations reacted to this prediction, and thanks to their conservation efforts, the leopard's population was stabilized to the extent that in 2017, the International Union for Conservation of Nature (IUCN) declared that this species is no longer an endangered species (McCarthy et al., 2017). This seems to me a clear case of SDR so that again, reactivity occurs when the theory at stake is taken from the *natural sciences* – from *zoology*.[20]

If so, it is clear that reactivity (so defined) is *not* exclusive to the human/social sciences – that ET1 is *false*. This is because agents can react in a causally effective way to theories' predictions from *many* scientific disciplines, from both human/social *and* natural sciences, at least in some cases, so as to obtain condition 4 in our formulation. This is due to agents' ability to react effectively also when they are not the target of the prediction, or, as Lowe (2021) puts it, "content-responsive actions [do not have to be] *on the part of the target system itself*" (p. 102).

Let us now turn to examine ET2. Here, let us begin by recalling our formulation of reactivity – reactivity can occur only where *agents* are involved. This is because it seems rather plausible that agents and only agents can *react to their beliefs*. Now, the question boils down to whether or not only human beings/social actors *are*, or *can be*, agents. I think that putting the question this way already takes us half the way to the answer.

Let us now turn to evaluate Grünbaum and Buck's debate. Grünbaum (1956) suggested the following example:

> … [C]onsider the goal-directed behavior of a servo-mechanism like a homing device which employs a feed-back and is subject to automatic fire control. Clearly every phase of the operation of such a device constitutes an exemplification of one or more *physical* principles. Yet the following situation is *allowed* by these very principles: a computer predicts that, in its present course, the missile will miss its target, and the communication of this information to the missile in the form of a new set of instructions induces it to alter its course and thereby to reach its target, contrary to the computer's original prediction. How does this differ, in principle, from the case where the government economist's forecast of an oversupply of wheat has the effect of instructing the wheat growers to alter their original planting intentions? (pp. 239–240)

It seems that if we accept this example as a proper case of reactivity, ET2 is false. However, in response, Buck (1963a) argues that Grünbaum's example is inadequate as the machine in the example does not act on *beliefs*, but on *orders* – it receives a set of *instructions*. Since obeying orders is different from acting on beliefs and so-called reflexivity requires the latter, Buck holds, Grünbaum's example is inadequate (pp. 366–368). But then, Grünbaum (1963) replies to Buck and states that his counterex-

[20] I refer the reader to an interesting example of Lowe's (2021) self-fulfilling science in the similar subject matter of animal speciation (pp. 115–118).

ample aims at Merton's (1949) account of so-called suicidal prophecies and not at Buck's (1963a). Subsequently, Buck (1963b) concludes that in this alleged debate, they were "engaged in a polite exercise of talking past each other" (p. 373).

Buck could not have put it more accurately. In this discussion, one's formulation of reflexivity/reactivity *already pre-determines* whether or not ET2 is correct. As Grünbaum (1956) and Buck (1963a) employ different formulations, their alleged debate is actually futile. Vetterling (1976) observes this fact and mentions that Buck's formulation pre-restricts the validity of reflexivity to cases in which *social actors* are involved (pp. 280–281). Indeed, in Buck's (1963a) formulation, ET2 is *ipso facto* true since the involvement of social actors is considered necessary in Condition 2 (p. 362).

What can be learned from the alleged Grünbaum and Buck's debate is that the way in which we formulate reactivity already takes us half the way to the answer. Buck's formulation is too narrow in this sense as at least *de jure*, not only social actors can react to beliefs, but also rational aliens, non-human animals, sophisticated robots, or advanced artificial intelligence systems. Hence, the generality of my formulation here seems more adequate.

Now, let us turn back to the question sketched above: Whether or not only human beings/social actors *are*, or *can be*, agents. In fact, this question is distinguishable into two distinct questions: (1) whether only human beings/social actors *are*, *de facto*, agents; (2) whether only human beings/social actors *can be*, *de jure*, agents. It seems surprisingly easy to respond to these questions without resorting to any examples. As to question (1), it seems that to the best of our knowledge, nowadays, human beings are the only agents known to us. No other known entity reacts to beliefs to obtain its goals.[21] However, as to question (2), it seems that there is no reason to assume that only human beings can possibly be agents. As I have mentioned here, it is rather possible that rational aliens, non-human animals, sophisticated robots, or advanced artificial intelligence systems can react to beliefs to obtain their goals. I suggest, then, that our judgment of ET2 should be as follows: It seems *de facto* true and *de jure* false.

To sum up, in this paper, I formulated reactivity (Sect. 2), examined the implications of self-defeating reactivity for predicting (Sect. 3), and rethought the exclusivity theses (Sect. 4). My point of departure was the fact that theory-deduced predictions might change agents' beliefs, and thus also agents' behavior. Since agents react to their beliefs by modifying their behavior to obtain their goals, they might react to a belief inspired by a theory-deduced prediction by modifying their behavior to obtain their goals, and this may have implications for the theory and its predictive success. What I regarded as the essence of this phenomenon is the agents' *reaction* to the belief inspired by some theory-deduced prediction. I concluded that self-defeating reactivity *does* make it impossible to predict, at least in some cases, and that while

[21] It is worth noting that there is a hot debate regarding whether artificial intelligence systems have agency: While some hold that such systems do not have intrinsic, non-human dependent agency (see, for example, Popa, 2021), others generously ascribe agency to them (see, for example, Floridi, 2023). This paper takes a cautiousposition as to this question, but considering the wide impact and rapid development of artificial intelligence, future research should examine whether (and how) artificial intelligence may have agency in a sense relevant to reactivity.

reactivity is not exclusive to the human/social sciences, it is exclusive to cases where agents are involved. Thus, it is exclusive to cases where human beings/social actors are involved only *de facto*. Reactivity, and specifically self-*defeating* reactivity, if I am correct, seems a severe epistemic challenge to science. Precisely when a theory becomes public, acknowledged, and credible, it may also come to derive incorrect predictions, which may significantly decrease its predictive success.

Acknowledgements I am indebted to my supervisor, Alexander Prescott-Couch, as well as to Jean Baccelli and Uskali Mäki for helpful discussions, two anonymous reviewers for many helpful comments, and Nehama Baker for proofreading.

Authors' contributions Not applicable.

Funding Funding from the Rhodes Scholarship, Merton College, and the Faculty of Philosophy at the University of Oxford is gratefully acknowledged.

Data availability Not applicable.

Declarations

Competing interests The author has no relevant financial or non-financial interests to disclose.

References

Aubert, K. E. (1982). Accurate predictions and fixed point theorems. *Social Science Information, 21*(3), 323–348. https://doi.org/10.1177/053901882021003001

Beinhocker, E. D. (2013). Reflexivity, complexity, and the nature of social science. *Journal of Economic Methodology, 20*(4), 330–342. https://doi.org/10.1080/1350178X.2013.859403

Buck, R. C. (1963a). Reflexive predictions. *Philosophy of Science, 30*(4), 359–369. https://doi.org/10.1086/287955

Buck, R. C. (1963b). Rejoinder to Grünbaum. *Philosophy of Science, 30*(4), 373–374. https://doi.org/10.1086/287957

Csonka, T. (2024, September 12). *Hungary's food price cap violated EU law, CJEU rules*. bne IntelliNews. https://www.intellinews.com/hungary-s-food-price-cap-violated-eu-law-cjeu-rules-343133/

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *The American Journal of Sociology, 113*(1), 1–40. https://doi.org/10.1086/517897

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36*(1), 15. https://doi.org/10.1007/s13347-023-00621-y

Frey, B. S. (2018). The crucial role of reactivity in economic science. In P. Róna & L. Zsolnai (Eds.), *Economic objects and the objects of economics* (pp. 141–150). Springer.

Friedman, M. (1953). The methodology of positive economics. In *Essays in positive economics* (pp. 3–43). University of Chicago Press.
Grünbaum, A. (1956). Historical determinism, social activism, and predictions in the social sciences. *The British Journal for the Philosophy of Science, 7*(27), 236–240. https://doi.org/10.1093/bjps/VII.27.236
Grünbaum, A. (1963). Comments on professor Roger Buck's paper "Reflexive Predictions." *Philosophy of Science, 30*(4), 370–372. https://doi.org/10.1086/287956
Grunberg, E., & Modigliani, F. (1954). The predictability of social events. *The Journal of Political Economy, 62*(6), 465–478. https://doi.org/10.1086/257604
Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–383). Clarendon Press.
Hacking, I. (1999). *The social construction of what?* Harvard University Press.
Henshel, R. L. (1995). The Grünberg/Modigliani and Simon possibility theorem: A social psychological critique. *The Journal of Socio-Economics, 24*(3), 501–520. https://doi.org/10.1016/1053-5357(95)90020-9
Jackson, R. (1979). Snow leopards in Nepal. *Oryx, 15*(2), 191–195. https://doi.org/10.1017/S0030605300024327
Jiménez-Buedo, M. (2021). Reactivity in social scientific experiments: What is it and how is it different (and worse) than a placebo effect? *European Journal for Philosophy of Science*. https://doi.org/10.1007/s13194-021-00350-z
Khalidi, M. A. (2010). Interactive kinds. *The British Journal for the Philosophy of Science, 61*(2), 335–360. https://doi.org/10.1093/bjps/axp042
Kopec, M. (2011). A more fulfilling (and frustrating) take on reflexive predictions. *Philosophy of Science, 78*(5), 1249–1259. https://doi.org/10.1086/662266
Laimann, J. (2020). Capricious kinds. *The British Journal for the Philosophy of Science, 71*(3), 1043–1068. https://doi.org/10.1093/bjps/axy024
Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press.
Lowe, C. (2018). The significance of self-fulfilling science. *Philosophy of the Social Sciences, 48*(4), 343–363. https://doi.org/10.1177/0048393118767087
Lowe, C. (2021). *Self-fulfilling science*. De Gruyter.
MacFarlane, J. (2003). Future contingents and relative truth. *The Philosophical Quarterly, 53*(212), 321–336. https://doi.org/10.1111/1467-9213.00315
MacKenzie, D. A. (2006). *An engine, not a camera: How financial models shape markets*. MIT Press.
Mäki, U. (2013). Performativity: Saving Austin from MacKenzie. In V. Karakostas & D. Dieks (Eds.), *EPSA11 perspectives and foundational problems in philosophy of science* (pp. 443–453). Springer.
Marchionni, C., Zahle, J., & Godman, M. (2024). Reactivity in the human sciences. *European Journal for Philosophy of Science*. https://doi.org/10.1007/s13194-024-00571-y
McCarthy, T., Mallon, D. P., Jackson, R., Zahler, P., & McCarthy, K. (2017). Panthera uncia. *The IUCN Red List of Threatened Species, 2017*, e.T22732A50664030. https://doi.org/10.2305/IUCN.UK.2017-2.RLTS.T22732A50664030.en
Merton, R. K. (1949). *Social theory and social structure: Toward the codification of theory and research*. Free Press.
Mitchell, D. J. B., & Azevedo, R. E. (1975). Price controls and shortages: A note. *The Journal of Business, 48*(4), 571–574. https://www.jstor.org/stable/2352324
Munson, L., & Worley, M. B. (1991). Veno-occlusive disease in snow leopards (*Panthera uncia*) from zoological parks. *Veterinary Pathology, 28*(1), 37–45. https://doi.org/10.1177/030098589102800106
Northcott, R. (2022). Reflexivity and fragility. *European Journal for Philosophy of Science, 12*(3), 43–43. https://doi.org/10.1007/s13194-022-00474-w
Øfsti, A., & Østerberg, D. (1982). Self-defeating predictions and the fixed-point theorem: A refutation. *Inquiry, 25*(3), 331–352. https://doi.org/10.1080/00201748208601971
Popa, E. (2021). Human goals are constitutive of agency in artificial intelligence (AI). *Philosophy & Technology, 34*(4), 1731–1750. https://doi.org/10.1007/s13347-021-00483-2
Romanos, G. D. (1973). Reflexive predictions. *Philosophy of Science, 40*(1), 97–109. https://doi.org/10.1086/288499
Runhardt, R. W. (2021). Reactivity in measuring depression. *European Journal for Philosophy of Science*. https://doi.org/10.1007/s13194-021-00395-0

Schaller, G. B., Junrang, R., & Mingjiang, Q. (1988). Status of the snow leopard *Panthera uncia* in Qinghai and Gansu Provinces, China. *Biological Conservation, 45*(3), 179–194. https://doi.org/10.1016/0006-3207(88)90138-3

Simon, H. A. (1954). Bandwagon and underdog effects and the possibility of election predictions. *Public Opinion Quarterly, 18*(3), 245–253. https://doi.org/10.1086/266513

Sober, E. (1999). Instrumentalism revisited. *Crítica: Revista Hispanoamericana de Filosofía, 31*(91), 3–39. https://doi.org/10.22201/iifs.18704905e.1999.751

Soros, G. (2013). Fallibility, reflexivity, and the human uncertainty principle. *The Journal of Economic Methodology, 20*(4), 309–329. https://doi.org/10.1080/1350178X.2013.859415

Spike, J. (2025, March 11). *Hungary's leader orders price controls on basic foods as inflation spikes*. AP News. https://apnews.com/article/hungary-orban-price-controls-food-inflation-economy-d023ade0d2ea7d6eda8044e79cde1005

Tsou, J. Y. (2007). Hacking on the looping effects of psychiatric classifications: What is an interactive and indifferent kind? *International Studies in the Philosophy of Science, 21*(3), 329–344. https://doi.org/10.1080/02698590701589601

Van Basshuysen, P., White, L., Khosrowi, D., & Frisch, M. (2021). Three ways in which pandemic models may perform a pandemic. *Erasmus Journal for Philosophy and Economics, 14*(1), 110–127. https://doi.org/10.23941/ejpe.v14i1.582

Vetterling, M. K. (1976). More on reflexive predictions. *Philosophy of Science, 43*(2), 278–282. https://doi.org/10.1086/288681