Epistemological status of rationality principles in the social sciences: a structural invariance criterion

Jeremy Attard*

Abstract

In the social sciences, within the explanatory paradigm of structural individualism, a theory of action – like rational choice theory – models how individuals behave and interact at the micro level in order to explain macro observations as the aggregation of these individuals actions. A central epistemological issue is that such theoretical models are stuck in a dilemma between falsity of their basic assumptions and triviality of their explanation. On the one hand, models which have a great empirical success often rest on unrealistic or even knowingly false assumptions; on the other hand, more complex models, with additional more realistic hypotheses, can (trivially) adapt to a wide range of situations and thus loose their explanatory power. Our purpose here is epistemological and consists in wondering to which extent demanding realistic assumptions in such cases is a relevant criterion with respect to the acceptance of a given explanatory model. Via an analogical reasoning with physics, we argue that this criterion seems too strong and actually irrelevant. General physical principles are not just idealized or unrealistic, they can also be formulated in many different yet equivalent ways which do not imply the same fundamental unobservable entities or phenomena. However, the *classification of phenomena* that such principles allow to highlight does not depend, at the end, on any particular formulation of these basic assumptions. This suggests that some hypotheses in theoretical models are actually not genuine empirical statements that could be independently tested but only substrates of modeling embodying a classification principle. Thus, we develop a *structural invariance* criterion that we then apply to rational choice models in the social sciences. We argue that this criterion allows to escape from the epistemological dilemma without condemning formal approaches like rational choice theory for their lack of realisticness nor being stuck to any antirealist viewpoint.

Keywords: explanatory models, rational choice theory, epistemological criterion.

*Philosophy and History of Science department, University of Mons (Mons, Belgium) & Department of Sciences, Philosophies and Societes, University of Namur (Namur, Belgium).

1 Introduction

The current notion of rationality in the social sciences and humanities stems from different traditions: sociology, economics, probability and decision theory, game theory, or even more recently cognitive psychology. The most formally developed and discussed theoretical framework is known under the name of *rational choice theory* (RCT), in which individuals are assumed to choose among a set of possible choices according to an *expected utility* they associate to each of them – for instance, by maximizing it. RCT has nurtured all these traditions and is common, in one form or the other, to all of them. All along its development and refinement, it has been extensively used and exported in different disciplines as sociology [Olson, 1971], criminology [Becker, 1968, Wikström and Kroneberg, 2022], or political sciences [Collier and Hoeffler, 2004].[1] In addition to its empirical success, it is also often presented as epistemologically promising, in the sense that it offers to social sciences and humanities a promising unifying framework allowing them to be understood within the same epistemological paradigm as other scientific disciplines.

However, RCT has well-known empirical and epistemological limitations and drawbacks too. The most discussed criticism is that rational choice models often rest on basic assumptions about human behavior which are seen as unrealistic, highly idealized or even knowingly false. The topic is wide and for the sake of brevity, in this paper we restrict to *social* sciences' models. That is to say, models which aim at explaining social (i.e. macro) regularities and *not* individuals' behaviors, and for which the effect of social structures and social interactions are implemented. Thus, we take RCT as a working example of a theory of action used in models of social phenomena, and our discussion does not apply to psychology, decision theory or so, for which *explananda* are individuals' actions.

The central epistemological question is the following: does the unrealisticness of basic assumptions (i.e. at the micro level) irrevocably undermine the epistemological value of rational choice models aiming at explaining social (i.e. macro) facts? The classical dilemma in which RCT is trapped, without being specific to RCT, is the following: on the one hand, a good explanation derives much from as few hypotheses as possible, and thus the latter are necessarily unrealistic – and thus how could they claim to be explanatory? On the other hand, more realistic models necessarily rest on more hypotheses, under less constraints, and thus the resulting explanation becomes less falsifiable and thus more trivial.

Classical responses to this dilemma are either the development of alternative theories of action or the defense of an antirealist or instrumentalist epistemological paradigm in order to turn rational choices models into epistemologically acceptable ones.

Our proposal in this paper is an epistemological criterion which aims at escaping from this dilemma without condemning formal approaches like RCT for their lack of realisticness nor being stuck to a mere antirealist paradigm. Via an epistemological

[1] The references given are just illustrative, and not necessarily representative of the wide literature in each of these fields.

comparison with physical theories, we argue that demanding realistic basic assumptions is too strong a criterion which is not even reached in physics, without undermining the epistemological value of its models. Basic assumptions like general physical principles are not only idealized or unrealistic, they can also be formulated in many different – yet equivalent – ways which do *not* imply the same fundamental entities or mechanisms, but still support the same explanatory power. Yet, the classification of phenomena that such principles allow to highlight does not depend, at the end, on any particular formulation of these basic assumptions. Our approach rests on the idea that some basic assumptions in theoretical models are actually not genuine empirical statements (and thus the question of their realisticness is just irrelevant) but only ways of representing or embodying a classification principle. Thus, we develop a *structural invariance* epistemological criterion that we then apply to RCT in the social sciences.

The remainder of this paper goes as follows. In section 2, we present a general formulation of RCT (2.1) and how social structures and interaction can be naturally implemented within this framework, with concrete working examples used all along the paper. Section 2.2 is then dedicated to the presentation of the epistemological issues RCT faces. We then make in section 3 a short interlude in order to distinguish between different kinds of basic hypotheses, clarifying some points – notably: what kinds of hypotheses are required to be realistic, and in which sense? This clarification allows us to reformulate the problem we tackle in this paper. In section 4, we translate this epistemological issue in the domain of physics, then develop a structural invariance argument from these reflections. Finally, in section 5 we go back to social sciences, develop this structural invariance argument and see how this leads to an epistemological criterion which aims at solving the dilemma we start from. In conclusion, we address some possible criticisms of our approach, and argue that our criterion, while accepting unrealisticness of basic assumption as unproblematic, does not necessarily commit us to any form of antirealism in the philosophical sense of the term.

2 Rational Choice Theory in the Social Sciences

2.1 General presentation

The general paradigm of explanations in the social sciences in which our work takes place, and within which the realisticness of RCT assumptions are examined, is a particular form of methodological individualism, namely *structural individualism* [Wippler, 1978], usually represented by the Boudon-Coleman diagram [Boudon, 1986, chapter 2], [Coleman, 1990, chapter 1], [Ylikoski, 2021] relating macro and micro levels of analysis, as sketched in figure 1.

As an illustration of structural individualism with RCT as a theory of action at the micro-level, we consider Breen and Goldthorpe's (BG) model of educational choices [Breen and Goldthorpe, 1997] as recasted by [Tutić, 2017], in direct line with Raymond Boudon's seminal work [Boudon and Lipset, 1974] applying RCT to social reproduction and educational system. In the BG model, the agents can originate from three possible

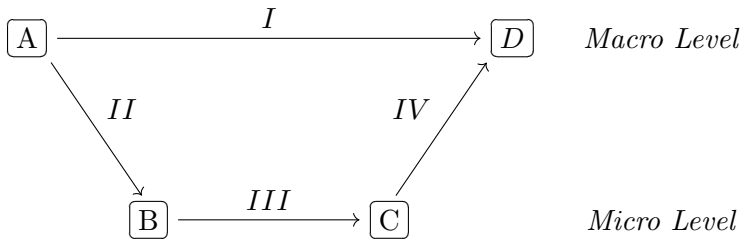

Figure 1: The Boudon-Coleman diagram

social classes: service class (S), working class (W) or under class (U). The main goal of this model is to explain that while costs of education had drastically decreased over time, leading to the increase of the proportion of children staying in the school system (that we denote here $P(stay|x)$ for all social classes $x \in \{S, W, U\}$), educational differentials between classes remained high, e.g. measured by odd-ratios such that:

$$\frac{P(stay|S)}{1-P(stay|S)} \Big/ \frac{P(stay|W)}{1-P(stay|W)} > 1. \tag{1}$$

In figure 1, this intringuing fact to be explained is represented by arrow I, which relates macro state A to macro state D, usually under the form of a causality relationship – here, a postulated causal influence of social origin (A) upon educational choices (D). According to structural individualism, explaining this macro-level observation means deriving it from micro-social considerations in three steps. Arrow II relates macro state A with micro state B. The latter could be the set of possible choices a student can face (for instance, as in the BG model, *leaving* the educational system or *staying* in it) and arrow II represents how the original class of the student affects their possible choices. Arrow III relates micro state B to micro state C by specifying, among possible choices of the agent, the one which is chosen and under which conditions. Finally, arrow IV represents how micro decisions aggregate into the macro state D: depending on actual choices agents do make, it allows to reconstruct the macro pattern which called for an explanation on the first place.[2]

Rational choice theory (RCT) is a possible theory of action to be used at the micro-level, represented by arrow III. Agents are assumed to face a set $\mathcal{A}$ of possible actions[3] leading to a set of given outcomes Ω with some probabilities. That is to say, $\omega \in \Omega$ is a possible outcome of any action $a \in \mathcal{A}$ with probability $p_a(\omega) \in [0, 1]$ to occur while doing action a. For example, the BG model focuses on the choice of classes S and W agents. Outcomes are the different social classes: $\Omega = \{U, S, W\}$ students can end up in, while possible actions are $\mathcal{A} = \{\text{staying}, \text{leaving}\}$. It they stay, they can succeed with

[2]Notice that since the macro observation bears on the actions chosen by the agents (stay or leave the educational system) and not outcomes of these actions (reaching social classes U, W or S), the arrow IV is trivial in the case of BG model.

[3]They can also be called choices, preferences, lotteries, etc. depending on the scientific context in which the theory is formulated.

probability π or fail with probability $1-\pi$. The possible outcomes of this binary choice and their corresponding probabilities are represented as a tree of choices, figure 2.

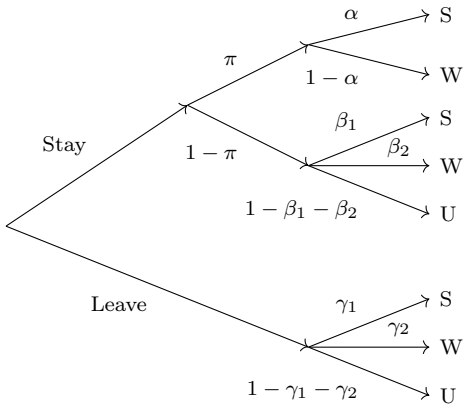

Figure 2: Schema of possible trajectories in the original Goldthorpe and Green's model.

For example, $(1-\pi)\beta_1$ is the probability to end up in class S having chosen to stay but having failed, and so on. Moreover, parameters $\alpha, \beta_1, \beta_2, \gamma_1$, and γ_2 are assumed to satisfy some relationships (see [Breen and Goldthorpe, 1997, p. 282] for more details) to reflect how much likely it is to end up in such or such social class following such or such trajectory.

The core of RCT is the assumption that each agent attaches a certain utility $u(\omega) \in \mathbb{R}$ to each outcome ω, and that the agent's actual choice is driven by the *expected utility*:

$$\mathcal{U}(a) = \sum_{\omega \in \Omega} p_a(\omega) u(\omega), \tag{2}$$

defined for each action $a \in \mathcal{A}$. For example, as it is often postulated, agents can be modeled as *utility maximizers*, i.e. choosing the action $a \in \mathcal{A}$ such that:

$$\mathcal{U}(a) = \max_{a' \in \mathcal{A}} \mathcal{U}(a'). \tag{3}$$

Modeling educational choices in the BG model then necessitates to associate to each action (stay or leave) a certain *expected utility*, which thus depends both on the utility attached by the agent to reaching classes U, W and S and on the different probabilities which structure the tree of choices pictured on figure 2. As noticed in [Tutić, 2017, p. 402], in the BG model the utilities attached to different outcomes are implicitly chosen as follows:

$$u_S(S) = \ u_W(S) = 1 \text{ and } u_S(U) = \ u_W(U) = 0 \tag{4}$$

$$u_S(W) = 0 \text{ and } u_W(W) = 1. \tag{5}$$

where $u_x(y)$ is the utility attached by an agent from social class $x \in \{\mathrm{S}, \mathrm{W}\}$ to reaching social class $y \in \{\mathrm{S}, \mathrm{W}, \mathrm{U}\}$. Hypothesis (5) is a way of implementing *relative*

risk aversion (families are assumed to try to avoid downward social mobility), which is the central mechanism in the BG model. Expected utility $\mathcal{U}_S$ ($\mathcal{U}_W$) for agents from social class S (W) thus reads:

$$\mathcal{U}_S(stay) = \pi\alpha + (1-\pi)\beta_1,\ \mathcal{U}_S(leave) = \gamma_1$$
$$\mathcal{U}_W(stay) = \pi + (1-\pi)(\beta_1+\beta_2),\ \mathcal{U}_W(leave) = \gamma_1+\gamma_2$$

From expected utility maximization principle, an agent from social class S stays in the educational system if:

$$\pi\alpha + (1-\pi)\beta_1 > \gamma_1, \tag{6}$$

and an agent from social class W stays if:

$$\pi + (1-\pi)(\beta_1+\beta_2) > \gamma_1+\gamma_2, \tag{7}$$

and so on. As illustrated with the BG model,[4] in this paper we restrict our analysis to the use of RCT in cases for which *social* (i.e. macro) facts have to be explained, and not decisions or behaviors of single individuals. Of course, even in these cases RCT is used to model individuals' decisions. However, in this context it is not an end in itself, as e.g. in decision theory, but merely a means to another end (deriving macro social phenomena). This distinction is fundamental for our argument, and we agree with [Goldthorpe, 1998] who argues that the main objective of a social sciences' model is to explain macro phenomena and not individuals behaviors, as it is also defended in [Hechter and Kanazawa, 2019, p. 3]:

> [There is a] common misconception about the nature of rational choice. The theory does not aim to explain what a rational person will do in a particular situation. That question lies firmly in the domain of decision theory. Genuine rational choice theories, by contrast, are concerned exclusively with social rather than individual outcomes.

The BG model is also an illustration of how social structures can be quite easily implemented in this kind of models. Here, arrow II makes the job: utility attached to each possible outcome (and thus to each possible action) differ between W and S classes, to take into account the central mechanism of this model, namely relative risk aversion. More generally, the effect of social structures (for instance, access to economical resources or social network aspects) in which agents are embedded can be straightforwardly modeled in this framework, being encoded in the expected utility attached to each possible action which can differ with respect to the macro state under consideration. Social interaction, that is interactions between individuals, can also be easily implemented using e.g. game theoretical framework as the influence of others of one's choice. These

[4] [Tutić, 2017] or [Becker, 2022] provide recent reviews of the BG model, its possible extensions as well as its theoretical and empirical successes or drawbacks.

examples show that structural individualism and RCT are wide enough frameworks not to restrict to atomic individuals acting out of any social structure or peer's influence, so that this straw man criticism can already be discarded.

2.2 Epistemological aspects and criticisms

Criticisms of RCT in general and applied to social sciences in particular are structured along a tension between two equally unsatisfactory situations: unrealisticness of basic assumptions and triviality of explanations.

On the one hand, and despite great empirical success, basic assumptions of RCT are often seen as unrealistic with respect to what is known about human behavior.[5] For example, as Margaret Mooney Marini formulated it in [Coleman and Farraro, 1992, pp. 24-25]:

> The most obvious problem with the axiomatization of utility and probability as a theory of the way people behave in choice situations is that it assumes that people have a high level of knowledge and computational ability with which to determine and evaluate a set of available alternatives. It assumes knowledge of all the alternatives available, as well as the consequences that will follow from each of the alternatives.

This debate has taken a particular form over the last thirty years in the wake of the development of *analytical sociology*, currently the most salient representative of structural individualism. Within this field of research, the position with respect to theories of action "progressively moved toward a more and more explicit pluralistic claim" [Manzo, 2021, p. 32] – that is to say, stopped considering RCT as the only (or even as a good or fruitful) possible starting point for the description of micro behaviors. Peter Hedström, for instance, developed his "desires-beliefs-opportunities" (DBO) theory in order to palliate some of these drawbacks which seemed to lead to an epistemologically unacceptable instrumentalism [Hedstrom, 2005]. According to him, action theories have to be, among others things, "psychologically plausible" [Hedstrom, 2005, p. 35] in the sense of being consistent with what is known about human behavior, e.g. in psychology, for models of social phenomena to be acceptable. He does acknowledge that our scientific models need to rest upon assumptions which somehow approximate reality. Yet, according to him, there is a crucial difference between approximations and false assumptions. He summarizes his concerns about "an unfortunate instrumentalist tendency": "Knowingly accepting false assumptions because they lead to better predictions or to more elegant models threatens the explanatory value and the long-term viability of the rational-choice approach." [Hedstrom, 2005, p. 9].

Since then, Hedström became even "considerably more skeptical" [Hedström, 2021, part 4] about the necessity and viability to implement individual intentions into models of micro-mechanisms. This implementation, to produce a genuine explanation, "has to be

[5] Whole fields of research, like e.g. behavioral economics [Page, 2022], even constituted upon the accumulating empirical observations [DellaVigna, 2009] that axioms of RCT are systematically violated.

firmly anchored in known facts about the acting individuals and their social settings." [Hedström, 2021, p. 498]. That is to say, explaining a macro fact means reducing it strictly to *observable* features of individuals' behaviors, and not to abstract models about what is supposed to occur in their minds – unless this information is available to the social scientist.

In addition to DBO, other theories of action, as Kroneberg's selection frame theory [Manzo, 2014, Chapter 4] or Franz Dietrich and Christian List's reason-based choice [Dietrich and List, 2013, 2016] developed over the last decade.

On the other hand, it is possible to stay within the framework of RCT and to complexify the basic hypotheses to make them more realistic, in line with *bounded rationality* approaches to human behavior [Simon, 1957, Wheeler, 2020]. That is to say, enlarging the definition of rationality to account for observed deviations. A possibility is to replace strict maximization principle (3) by a probabilistic principle as described in [Kruis et al., 2020], used e.g. in Manzo's model of educational choices with social interaction [Manzo, 2013a, p. 58], in the same vein as what is developed in game theory under the name of *quantal response equilibrium* [McKelvey and Palfrey, 1995]: every agent has probability $P(a)$ to do action $a \in \mathcal{A}$, with:

$$P(a) = \frac{e^{\beta U(a)}}{\sum_{a' \in \mathcal{A}} e^{\beta U(a')}}, \tag{8}$$

for $a \in \mathcal{A}$ and $\beta > 0$, in the case for which $\mathcal{A}$ is a finite set (the generalization is straightforward). (8) reduces to (3) for $\beta \to \infty$ and thus is a generalization of the latter, and P is uniform for $\beta = 0$; that is to say, in the latter case the agent acts purely "randomly".

Prospect theory [Tversky and Kahneman, 1992, Barberis, 2013] is another example of a complexification of these basic hypotheses. In its simplest form, it consists in modifying probabilities p_a over the set of possible outcomes Ω of an action $a \in \mathcal{A}$ to take into account *biased subjective probabilities*, setting a reference point and giving utility function a S-shape which "is concave above the reference point and convex below it but less steep above than below" [Coleman and Farraro, 1992, Chapter 2, p. 26]. This particular shape models reference-dependence in preferences, diminishing returns and loss aversion, as observed in psychology and behavioral economics [Page, 2022].

Notice that these are just few possibilities among others, possibly infinite shapes for these "risky curves" [Friedman, 2017]. It is e.g. also possible to take into account the temporal dimension of preferences (and more particularly apparent time inconsistency) by adding hyperbolic corrections to standard utility function [Gintis, 2009] or subjective beliefs within Bayesian decision theory [Binmore, 2011] "without modifying any of [RCT's] basic conceptual elements" [Manzo, 2013b, p. 369].

This approach aims at being more realistic but is also often criticized for several reasons. According to [Moscati, 2023], these so-called alternative models to "neo-classical theory" are nothing but (maybe more elaborated) as-if models, and thus do not save it from classical epistemological criticisms. However, the main criticism which interests us here is that these kinds of theoretical adjustments easily end up to *trivialize* the

underlying explanation. Indeed, the less constraints there are on the way the rationality principle concretely instantiates, the easiest it is to produce an empirically adequate model – but in this case, the explanation is trivial. For instance, quantal response equilibrium model in game theory, cited above, is criticized precisely for this reason [Haile et al., 2008]. Anyway, this case of a too loose RCT which trivially explains everything is actually a particular instance of what Karl Popper qualified as unfalsifiable theory [Popper, 1962], or what Imre Lakatos called a degenerating problemshift [Lakatos, 1978]. This issue has extensively been recognized in the social sciences. For example, [Kroneberg and Kalter, 2012, p. 83]:

> The wide version of RCT is able to assimilate almost any psychological concept or theory and translate it into more or less "soft" incentives or a more or less inaccurate belief.

or Hedström and Ylikoski in [Manzo, 2014, Chapter 2, pp. 59-60]:

> From the point of view of the core assumptions of RCT, broadening the range of individuals' concerns to non-monetary goods and to the welfare of others is straightforward, although there is concern about whether this can be done in a non-arbitrary manner. (...) Finding a RCT model that fits a particular phenomenon becomes almost trivially easy as there are no real constraints on preferences and beliefs that can be attributed to the individuals in question.

or in [Hedström, 2021, p. 498]:

> What Becker and his followers showed is that it is possible to come up with coherent rational narratives that fit even the most puzzling and seemingly irrational kinds of behavior, particularly if the narratives are free to ignore known empirical facts about the behavior of individuals.

The same kind of criticisms apply to reason-based models of human behavior, in the sociological tradition as Boudon's "generalized rational-choice model" [Boudon, 1996]. The latter is based on the idea that actors whose behaviors are aimed to be explained should be assumed to have *good reasons* to behave as they do, such that "these reasons can be in some circumstances of the cost-benefit comparison type, but in other circumstances of other types" [Boudon, 1996, p. 147]. The central criticism addressed to Boudon's "cognitivist model" is that it seems to be always possible, a posteriori, to build up reasons a given actor has to behave the way he does, undermining the explanatory power of such a framework if no more constraints are added. In the same vein, in [Hedström, 2021, p. 490] Hedström qualifies as "just-so stories" models of micro-behaviors resting on unobservable assumptions about mental states of individuals, and urges to avoid this kind of pseudo-explanatory models.

Thus, it turns out that RCT (and generally action theories, in the context of structural individualism) is in a critical epistemological position between as-if models, which may have a great empirical success but rest on unrealistic assumptions, and just-so stories, which rest on more descriptively precise hypotheses – either aiming at being more

realistic or which are just unobservable – but so that the explanation end up to be trivial. This dilemma has long been recognized, e.g. by Neil J. Smelser [Smelser, 1992] quoted in [Hernes, 1992, p. 421], which perfectly summarizes the situation:

> the postulates of rational choice are not realistic, as the assumptions (...) leave out key aspects of human behavior (...) [and] the responses to incorporate criticism of rational choice theory have led to a watering down and a movement in the direction of theoretical indeterminacy – more abstract, more truistic, more context free, incapable of falsification – which "produces the specter of an inclusive and universally applicable construct that simultaneously explains everything and therefore nothing."

This dilemma is also reflected in the opposition between "narrow" and "wide" version of RCT (the latter weaken rationality postulates with respect to the former) [Opp, 2013, Manzo, 2013b], or between "thin models" (which do not assume anything about individuals' motivations) and "thick models" (which do take into account more complex psychological features of individuals) of rational choice [Hechter and Kanazawa, 2019].

This paper focus on epistemological considerations, that is to say: how to *justify* the use of some hypotheses rather than others, while escaping from the dilemma described above. In other words, what criteria can we demand to be satisfied in order to consider some hypotheses as acceptable, beyond the mere empirical-success criterion which, for all reasons already cited, seems not to suffice.

3 Clarification interlude: what hypotheses are we talking about?

An important question to ask when it comes to talk about the realisticness of hypotheses is precisely: *what kinds of hypotheses are required to be realistic, and in which sense?* In this paper we focus on hypotheses which belong to theoretical (i.e. explanatory) models in science, and in particular in the social sciences within the framework of structural individualism using RCT as a theory of action. A first step is to highlight their general structure, and in particular the different types of hypotheses which compose them. We assume here that theoretical models can be described as being composed of three distinct types of hypotheses: general principles (or theoretical framework) $\mathcal{P}$, theoretical hypotheses $\mathcal{H}$ and initial conditions (or contextual assumptions) Γ. This view of theoretical models M as a triple $M = (\mathcal{P}, \mathcal{H}, \Gamma)$ is in line with the quite usual *hierarchical view* of models [Suppes, 1966, Giere, 2009, Winther, 2016]. The basic idea, in a semantic conception of theories, is that scientific knowledge is structured under the form of models of several levels of abstraction. This goes from the most concrete one, (most) directly connected to empirical realm, to the most abstract one, composed of general principles, or axioms. Since we focus on theoretical models, we do focus on the most abstract part of this hierarchical structure.

In this paper we use the following quite general pattern of explanation: an empirical fact F, to which some contextual assumptions (or initial conditions) Γ are attached,

is said to be explained within a theoretical framework $\mathcal{P}$, if there exists a finite set of theoretical hypotheses $\mathcal{H}$ such that F can be logically derived by the conjonction of $\mathcal{P}$ and $\mathcal{H}$ using Γ as correspondence rules between abstract entities in the model and concrete elements of the empirical fact to be explained. This view is very general and does not assume anything about the epistemological quality of the explanation studied. We represent this pattern of explanation schematically as:

$$\exists\ \mathcal{H} \mid \mathcal{P} \cdot \mathcal{H} \xrightarrow{\Gamma} F. \tag{9}$$

Let us first take an example in physics to illustrate this structure: the fall-with-friction case, where a spherical marble is dropped from an initial height h_0 in a given fluid (like air) with no initial speed. The trajectory of the marble (a material point of mass m) is represented as a function $t \longrightarrow \mathbf{r}(t) = (x(t), y(t), z(t)) \in \mathbb{R}^3$ which is assumed to satisfy the fundamental principle of dynamics:[6]

$$m\ddot{\mathbf{r}} = \sum_i \mathbf{F}_i. \tag{10}$$

where $\{\mathbf{F}_i\}_i$ is a finite set of forces which are assumed to apply to the marble. More precisely, in this case, we consider Earth's gravity force:

$$\mathbf{F}_1 = m\mathbf{g} \tag{11}$$

and friction force:

$$\mathbf{F}_2 = -\alpha\dot{\mathbf{r}}. \tag{12}$$

Equation 10 can then be solved, giving the trajectory $t \longrightarrow z(t)$ as a solution which can eventually be compared to data. In this case, the three kinds of hypotheses are the following:

General Principles $\mathcal{P}$: they consist in the general Newtonian framework together with equation (10).

Theoretical hypotheses $\mathcal{H}$: they flesh out the general principles by specifying the set $\{\mathbf{F}_i\}_i$. More precisely, there are two kinds of hypotheses here:

- which forces are considered: here there are Earth's gravity force and friction force. In free fall model, only gravity force is considered, the effect of fluid is neglected.
- what are their precise mathematical form: here they are given by equations (11) and (12). (11) is one way to represent Earth's gravity. Another one is the well-known and more general Newtonian gravitational force in $1/r^2$. Similarly, (12) is one way to represent the friction force. Another possibility sometimes encountered is a force proportional to speed squared.

[6] In the whole paper, vectors are represented in bold.

Initial conditions Γ: here they consist in the specification of the initial values of $\mathbf{r}(t)$ and its derivative $\dot{\mathbf{r}}(t)$: $\mathbf{r}(0) = (0, 0, h_0)$ and $\dot{\mathbf{r}}(0) = \mathbf{0}$. The independent measure of the mass m is also something which can count as such kind of hypotheses.

The case of a rational choice model exhibits the same kind of structure:

General Principles $\mathcal{P}$: they consist in the general framework of RCT as described in section 2.1 and a principle relating utility (2) to action, like (3) or (8).

Theoretical hypotheses $\mathcal{H}$: here again, these hypotheses concretely instantiate the general principles, and are divided in two types:

- what types of utility are considered: economical, social, cognitive, axiological, ... ? In the BG model as presented in section 2.1, only economical utility is considered but in the original model cognitive ability is also taken into account. In more elaborated models like [Manzo, 2013a], social interaction is also taken into account in addition to economical and cognitive aspects.
- what are their precise mathematical form ? In the case of the BG model, economical utility is reduced to a binary value (equations (4) and (5)). Taking into account cognitive ability means attributing to it a certain distribution across population (typically normal or log-normal). Taking into account social influence leads also to a certain utility term which has a certain form, for instance a quadratic form like in the Brock and Durlauf's model of theory choice in science [Brock and Durlauf, 1999].

Contextual assumptions Γ: they consist in the specification of the structure of choices or actions considered. For example, in the BG model, contextual hypotheses are the fact to consider only three social classes, and the whole structure of choice represented in figure 2, which are assumed to be empirically given.

The classical dilemma presented in sec. 2.2 can now be rephrased as follows: on the one hand, if no constraints are imposed to the set $\mathcal{H}$, then (9) is trivially true and the explanation is empty; on the other hand, a less trivial explanation thus means more constraints on the set $\mathcal{H}$ which then turn out to be oversimplified and thus unrealistic assumptions. The same argument applies to $\mathcal{P}$ itself: the more it is specific, the less (9) is trivial but the less $\mathcal{P}$ is realistic, and vice versa.

Contextual hypotheses/initial conditions Γ have a more direct empirical meaning and also serve to connect abstract entities in the model to concrete features of the system studied. In other words, their realisticness is less questionable, because by construction they are fitted to directly represent measurable variables. For example, it is obvious that if the initial position of a physical system can be measured and is equal to x_0, then this value has to be given to $x(0)$ in the model. In the social sciences, Γ are used to connect abstract models to concrete social or political situations which aim at being scientifically studied. For the BG model to be applied to a concrete academic system and a political situation of a particular country, the structure of choices of the BG model, for instance,

should be connected to actual possibilities offered to individuals. In this paper, from now on, we thus consider that Γ are given in all cases studied and *we focus on realisticness of* $\mathcal{P}$ *and* $\mathcal{H}$.

In the case of RCT, concerns about realisticness of these two kinds of hypotheses can be formulated as follows. On the one hand, we can wonder whether individuals are *really* computing expected utilities – trying, for instance, to maximize them. On the other hand, if we assume RCT basic assumptions, other questions could bear on the different types of utility considered: are they *really* the only ones in presence? Are their mathematical forms the *real* one? Our tripartite view of theoretical models is in part methodologically justified by the fact that these two kinds of questions are of distinct nature and do not call for the same kind of answers. In this paper, we focus mostly on the former, i.e. on general principles $\mathcal{P}$, asking e.g. whether the realisticness of the maximization principle in RCT is relevant with respect to its legitimacy to be used as an explanatory principle. We do not focus so much on the latter kind of questions. Indeed, they are less epistemologically challenging: if one agrees on the use of abstract and unrealistic general principles, then it is quite consensual that theoretical hypotheses $\mathcal{H}$ used in a given model, like the different kinds of utilities considered in a rational choice model or the different forces taken into account in a Newtonian mechanical model, are those which are assumed to play a relevant role in the derivation of the explanandum, and *not* all possible kinds of utility or forces actually in presence. Free fall model, for instance, when it is empirically successful, does not assume that there is no fluid (in this case it would be indeed a unrealistic assumption) but only that given the conditions in presence (described by contextual assumptions Γ), the presence of the fluid has no consequence on the outcome, and thus is not taken into account.

Before going on, let us precise one thing. "Realisticness" of hypotheses should not be strictly confused with "realism" in the sense of *scientific realism*, that is to say whether there exists an objective and independent reality (metaphysical realism), whether the content of a theory (dealing with its unobservable entities) should be taken literally (semantic realism) or whether we can or not deduce from the empirical success of our best theories that their postulated and unobservable entities exist (epistemic realism) [Psillos, 1999]. Even if these features are obviously related, when scholars talk about the "realisticness" of basic assumptions in theoretical models, they often mean their *consistency* with what is otherwise known (about human behavior for instance). For example, as already mentioned, Hedström insists on psychological plausibility [Hedstrom, 2005, p. 35] as a necessary (or, a least, highly demanded) epistemological criterion for a theoretical model to be acceptable as part of a genuine explanation.

In the remainder of this paper, we challenge the latter assumption, that is the necessity (from an epistemological viewpoint) for a genuine explanatory model in the social sciences to rest on assumptions (of $\mathcal{P}$ and $\mathcal{H}$ types, again: we are not talking about Γ) that are consistent with what is otherwise known about human behavior. We argue, from a comparison with physical theories, that this is not the right epistemological criterion – and that under this criterion not many physical models would be acceptable. We present a criterion, *structural invariance*, which seems to be a better one for epis-

temological purposes. We also briefly argue that this epistemological viewpoint does not commit us to any form or scientific realism or antirealism. That is to say: even if consistency of basic assumptions $\mathcal{P}$ and $\mathcal{H}$ is not seen as a primary epistemological criterion, we argue that it does not mean that we are commited to instrumentalism or even any form of antirealism.

4 Epistemological status of hypotheses: a detour through physics

Our aim in this section is to show that "realisticness", or external consistency, of fundamental principles $\mathcal{P}$ and theoretical hypotheses $\mathcal{H}$ is an knowingly *unecessary* epistemological criterion in physics, not only in the discovery context, but also from the justification viewpoint. What can be more directly compared to features of the empirical fact to be explained are contextual hypotheses Γ. Only in their case external (at least approximated) consistency is indeed an important epistemological criterion.

4.1 Some hypotheses are knowingly false, unrealistic or even theoretically inconsistent

First, it is a quite longstanding and fairly consensual observation in epistemology and philosophy of science [Frigg and Hartmann, 2020] that in physics some hypotheses are often knowingly false, or highly idealized and thus unrealistic, while others are even forbidden by the theory itself.

For instance, explaining the trajectories of planets in the solar system within Newtonian mechanics assumes to describe them as highly idealized objects, for example as material points. We "know" that planets are not points; yet, such a model, once it makes good falsifiable but empirically adequate predictions, is considered as providing a genuine scientific explanation. We could model planets as spheres, but for a wide range of observations this "more realistic" assumption will not have any effect on the empirical adequacy of the model, and thus on its explanatory power. Let us notice that material points not only do not exist, but are forbidden by Newtonian physics, for they assume an infinite energy density. However, somehow, this (quite trivial) observation does not really impinge on the building and selection of physical models.

Another example is kinetic theory of gas, from which can be derived e.g. ideal gas law relating pressure P, volume V, temperature T and amount of matter n of a gas in particular conditions:

$$PV = nRT, \tag{13}$$

where $R = 8,314$ J mol^{-1} K^{-1} is the ideal gas constant. Relationship (13) can be derived from micro considerations, modelizing the gas as a set of N identical particles (material points of mass m) confined in a volume V. The particles are assumed to interact only with the walls of the box they are confined in via elastic shocks, and not between them. Due to the number of particles (about $\approx 10^{23}$), a strict application of

Newtonian mechanics is impossible: a statistical approach is necessary. More precisely, speed of particles is assumed to be distributed along a certain statistical distribution (here, Maxwell's law). Then, the strategy is to compute pressure P as the macro effect of the shocks of particles on the walls by counting the number of particles of a given speed (up to an infinitesimal amount) which reach an infinitesimal piece of the wall in an infinitesimal interval of time. This finally leads to the total force which exerts per surface unit, and to recover relationship (13).

In this model, we first assume that molecules of the gas are material points which do not interact with each others whereas they probably experience *millions* of shocks per second. This is thus not a mere approximation, but a clearly false assumption. Second, their speed are assumed to be distributed a certain way, while this assumption is not directly testable. Moreover, we know otherwise, e.g. from quantum mechanics, that atoms and molecules are actually much more complicated objects, and absolutely not mere points or even complicated assembly of spheres. In quantum mechanics, fundamental systems are described not as material objects like spheres or so, but as wave functions only giving information about probabilities of different states in which the system can be.

Thus, basic hypotheses of these models are not only unverifiable, but probably completely wrong. However, a fundamental observation here is that in the case of gas kinetics, we do not need to model molecules as in quantum mechanics. Indeed, it does not weaken the explanatory power of the statistical model, while a "more realistic" model, that is a theoretical model resting on quantum mechanics-based assumptions rather that simpler objects as material points, would *not* necessary be a better model – if its aims is to derive ideal gas law. Of course, it is then possible to relate gas kinetics assumptions to deeper considerations from quantum mechanics (for instance, relating atoms-as-spheres assumptions to spherical harmonics model of atom in quantum mechanics), and this external consistency is overall a good epistemological feature. However, our point here is that this connection is not necessary for a theoretical model to be acceptable.

4.2 Physical theoretical frameworks enjoy pluralistic and ontologically incompatible formulations

Second, as it is presented e.g. in [Suppe, 2000] as an argument against some logical empiricists' positions, a given physical theory can usually be formulated in several distinct ways such that these formulations, even if theoretically and empirically equivalent, postulate *different* entities or mechanisms. A well-known example is Newtonian mechanics, which can be formulated postulating either a set of forces $\{\mathbf{F}_i\}_i$ which act on material points such that their trajectory is related to $\{\mathbf{F}_i\}_i$ by fundamental dynamics principle (10), a Lagrangian $\mathcal{L}$ associated to the material point and the physical situation, such that the material point follows the trajectory γ which minimizes the *Action*: $\int_\gamma \mathcal{L}dt$, or a Hamiltonian $\mathcal{H}$ driving the dynamics of the physical system, seen as a point $(\mathbf{p}, \mathbf{q})$ in a phase space, according to Hamilton equations: $\frac{dq}{dt} = \frac{\partial \mathcal{H}}{\partial p}$ and $\frac{dp}{dt} = -\frac{\partial \mathcal{H}}{\partial q}$. More specifically, classical *gravity* can be formulated in the Newtonian framework by postu-

lating well-known $1/r^2-$gravitational force, but can also be formulated in the Newton-Cartan geometrical framework without postulating forces but rather describing it as the manifestation of the curvature of an underlying (classical) spacetime [Ehlers, 1973, Chapter 1]. This is not a specificity of Newtonian mechanics, but a feature shared by all physical theories. For instance, quantum mechanics acknowledges at least nine different formulations [Styer et al., 2002]: Heisenberg's (matrix), Schrödinger's (wavefunction), Feynman's (path integral), Wigner's (phase space), density matrix, second quantization, variational, de Broglie-Bohm's (pilot wave) and Hamilton-Jacobi's. As for General Relativity, it can also be formulated different ways [Göckeler and Schücker, 2011, Krasnov, 2020, Arnowitt et al., 2008]: using tensor calculus on differential manifolds, Cartan geometry on principal fiber bundles, encoding gravitation in torsion instead of in curvature, or deriving fields equations from Lagrangian or even Hamiltonian principles.

All these formulations are equivalent in the sense that it is perfectly known how to mathematically pass from one to the other, and have the same empirical content. However, entities that they postulate (forces or lagrangians, quantum states as vectors or density matrix, spacetime torsion or curvature, and so on) together with fundamental mechanisms and laws which lie at the foundation of their explanatory power have sometimes nothing to do with each other. Picture of reality given by these formulations are thus deeply distinct, even though the latter are epistemologically equivalent and all highly acceptable as explanatory models.

Moreover, sometimes entities postulated by a theory T are no more postulated by a overcoming theory T', i.e. such that empirical success of T is strictly included in empirical success of T'. A well-known case is general relativity overcoming Newtonian gravitation theory. However, this fact never weakened the explanatory power of Newtonian physics. Thus, knowingly "false" or merely overcome general principles can still enjoy a fairly high epistemological value.

4.3 General principles and theoretical hypotheses are hardly independently testable

Third, more generally, it can be argued that general principles and theoretical hypotheses cannot actually be as directly tested as e.g. contextual assumptions can be. More precisely, and as already mentioned, general principles, without more precision, make the pattern of explanation (9) trivially satisfied: given an empirical fact, there always exists a set of theoretical hypotheses $\mathcal{H}$ (even if it means getting a little creative) such that it is possible to explain this empirical fact within the framework defined by $\mathcal{P}$. Let us consider fundamental principles of Newtonian dynamics. In this framework, explaining the trajectory $\mathbf{r}(t)$ of a system means finding a finite set of forces $\{\mathbf{F}_i\}_i$ such that (10) is satisfied, i.e. $m\ddot{\mathbf{r}} = \sum_i \mathbf{F}_i$. However, strictly speaking, *it is always possible*, given a certain trajectory $\mathbf{r}(t)$, to find a certain set $\{\mathbf{F}_i\}_i$ such that (10) holds – just as it is always possible, given a certain human behavior, to find a utility function over a set of choices such that this behavior corresponds to the maximization of this utility function. Moreover, each time a given model is apparently refuted, it is always possible to add or modify some hypotheses in order to make the model empirically adequate again.

The same observation holds for any general principle in physics: some terms (forces, a Lagrangian, a Hamiltonian, ...) are assumed to drive the dynamics of the phenomenon which aims at being explained within a set of general principles $\mathcal{P}$, but without more precision it is trivially always possible to find the right form for these terms to derive the observations from $\mathcal{P}$. Thus, wondering if hypotheses as $\mathcal{P}$ are "true" is not an empirical question.

As for theoretical hypotheses, it turns out that is is impossible to test them outside the framework defined by some general principles: $\mathcal{H}$ are always subordinated to some $\mathcal{P}$. For instance, it does not mean anything to wonder whether kinetic energy E_k of a material particle of mass m with speed v, in Newtonian mechanics, is *really* equal to $\frac{1}{2}mv^2$, because the hypothesis:

$$E_k = \frac{1}{2}mv^2 \tag{14}$$

does not have any meaning outside Newtonian mechanics, as well as e.g. the friction force (12). In other words, there is no way of measuring kinetic energy which could be compared independently to (14) in order to empirically test the latter.

Of course, general principles $\mathcal{P}$ can be sometimes derived as special cases of more general principles. For instance, classical equations for gravitation can be derived from Einstein's equations as a classical limit, i.e. for gravitation potentials Φ such that $\Phi \ll c^2$ where c is the speed of light. Theoretical hypotheses can also be derived as limit cases of more general hypotheses. For instance, Earth gravity force (11) is a special case of gravitational force $\mathbf{F} = -\frac{GmM}{r^2}\mathbf{e_r}$ for $M = M_T$ (Earth's mass) and $r = R_T + z$ with $z \ll R_T$, where R_T is the Earth's radius. However, in all these cases, the derivation also rests on other general principles or theoretical hypotheses – thus, epistemologically speaking, it only shifts the problem.

4.4 Outline of a structural invariance argument

Regarding the previous reflections, it turns out that theoretical models in physics cannot derive their epistemological value from empirical adequacy of their basic assumptions, because either the latter cannot be directly tested or, when they do and turn out to be false, unrealistic or overcome by other deeper ones, they do not seem to loose their epistemological value. Consistency of basic hypotheses in theoretical models in physics seems not only to be unnecessary but even sometimes impossible, without undermining their epistemological status. So, the question still remains: which criterion is at work for ensuring a theoretical model its justification?

Our argument goes as follows. Avoiding triviality for explanations (9) within a given theoretical framework $\mathcal{P}$ amounts to impose a set of *constraints* on the acceptable theoretical hypotheses $\mathcal{H}$. Looking at physical theories, our proposal is that relevant epistemological criteria actually do not apply directly to basic assumptions such as $\mathcal{P}$ or $\mathcal{H}$, demanding e.g. their accuracy, realisticness or external consistency, but rather to the set of constraints which are imposed to $\mathcal{H}$ in order to turn (9) into a non trivial explanation within $\mathcal{P}$. More precisely, what makes a theoretical hypothesis in physics (for example, gravitational term in Newtonian physics: $\mathbf{F} = -\frac{GmM}{r^2}\mathbf{e_r}$) epistemologically

acceptable is not that it is realistic or merely true (for reasons cited above), but rather that it is *systematically* associated to a certain set of empirical situations (here, some gravitational phenomena) with great empirical success (non trivial adequacy), and that in other formulations of classical mechanics it has a corresponding term which is also systematically associated to the same set of empirical situations with the same empirical success. Thus, the constraints imposed to the set of acceptable theoretical hypotheses ends up exhibiting a certain structure: a systematic association between classes of phenomena and classes of models which explain them. The final point is that even if general principles and theoretical hypotheses can be formulated in different ways, they provide isomorphic classifications of models and thus a classification of phenomena which does not depend on any particular formulation.

Therefore, *epistemologically satisfying general principles $\mathcal{P}$ are those principles which allow, from a finite and structured set of theoretical hypotheses, a salient classification of empirical facts by mapping them to a classification of models generated by $\mathcal{P}$ and a restricted set of theoretical hypotheses, such that the whole structure turns out not to depend on any particular formulation.* In the following section, we develop this argument on a more general basis. In particular, we argue that this observation is not restricted to physis and is rather a fundamental epistemological feature.

5 Neither *as-if* nor *just-so* stories: a structural invariance criterion

Our aim in this section is to show that it is possible to exhibit an epistemological criterion which does not condemn formal approaches like RCT while escaping from the dilemma described in section 2.2 without being stuck in a mere antirealist viewpoint. Our point is not a strict defense of RCT; we just take it as a working example because it is the most discussed theory of action in the literature, but our reflections aim at applying beyond it, i.e. to any theory of action used at the micro-level in the context of structural individualism. Our argument rests on an analogical reflection with physics made in section 4.

5.1 Generation of classes of models and classification

Once a set of principles $\mathcal{P}$ is given, a class of models can be generated by adding to it a specific set of theoretical hypotheses. For instance, fundamental principle of Newtonian dynamics $\mathcal{P}$ together with the hypothesis "$\mathbf{F} = -\frac{GmM}{r^2}\mathbf{e_r}$" generates the class of Newtonian gravitational models, which falls (free fall, fall with friction, etc.) constitute a subclass of. Still in Newtonian mechanics, another example is the class of conservative phenomena, defined by forces of the form: $\mathbf{F} = -\nabla U$, where U (the potential) is a scalar map. More generally, a class of models $\mathcal{M}$ is generated from a $\mathcal{P}$ by the addition of a specific finite set of theoretical hypotheses $\mathcal{H}$.

Let us consider a set of empirical facts F aiming at being explained within a framework defined by a set of general principles $\mathcal{P}$. Let T be a map which associates an

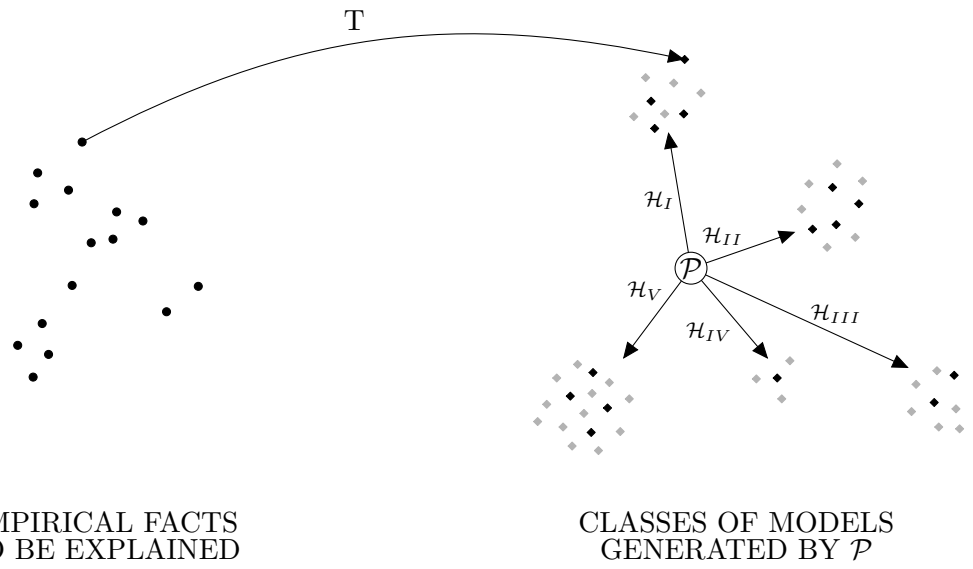

Figure 3: General principles as a cartographic map. Black filled circles represent empirical facts to be explained. Black filled squares represent theoretical models generated from $\mathcal{P}$ and a certain finite set of theoretical hypotheses, labeled as $\mathcal{H}_I$, $\mathcal{H}_{II}$, ... Gray filled squared represent models which are not (yet?) associated to any empirical fact but still belong to the space of models defined by $\mathcal{P}$. T maps empirical facts to theoretical models which explain them on a non trivial manner.

empirical fact and a theoretical model $M = (\mathcal{P}, \mathcal{H}, \Gamma)$ if the latter explains the former in the sense given above (9) and in a non-trivial manner. Figure 3 sketches this idea: a set of empirical facts are mapped to theoretical models which are generated from $\mathcal{P}$ by adding a specific set of theoretical hypotheses $\mathcal{H}$. Keep in mind that it is a dynamical process. That is to say, the constraints imposed on theoretical hypotheses are only partially a priori given (e.g. in an initial, pre-scientific classification of empirical facts) and they are in most part diachronically highlighted. In some epistemologically felicitous cases, a salient classification appears: a wide range of empirical facts are explained within a relatively small number of classes of models. Such a classification map is always possible to build, but does not necessarily give a non trivial and invariant classification. More precisely, our structural invariance criterion is given as follow:

> A set of general principles $\mathcal{P}$ derives its epistemological value from the extent to which it allows a salient (i.e. not trivial) classification of empirical facts within classes of models (corresponding to specific finite sets of theoretical hypotheses $\mathcal{H}$) which explain them via a map T as described above, such that this classification does not depend, at the end, on any particular formulation of $\mathcal{P}$ (or $\mathcal{H}$).

Let us now explore some consequences of this criterion, seeing how it applies to RCT assumptions in the social sciences.

5.2 Consequences of the criterion

The view presented here has some obvious connection with the fictional view of models [Frigg, 2009, Suárez, 2009], and more precisely with Alisa Bokulich model-based account of scientific explanation [Bokulich, 2017], for which the central question is how models containing inaccurate hypotheses can still provide genuine explanations of phenomena. The general answer given to this issue is that a scientific model, even if it rests on idealized or false assumptions, does explain as long as it satisfies some important conditions, like accounting for counter-factual patterns which characterize the phenomenon and being such that its target belongs to its domain of applicability. In this perspective, the actual adequacy of basic hypotheses of a model is no longer of primary importance in ensuring the model its epistemological value. From our viewpoint, the reflection above (section 5.1) is in the same line and suggests that we should not take fundamental principles or hypotheses too literally, as really describing what is going on beyond direct observations, but as a (necessary) representation of an underlying classification map T. "Realisticness" of these assumptions is thus not the relevant epistemological criterion at work: their role is only to concretely embody this classification, not to represent reality.

Confusion comes from that we confuse these kinds of hypotheses with initial conditions/contextual assumptions, which do have a more direct empirical meaning – and for which the independent measure is important, which can be approximated, etc. Indeed, some hypotheses or principles, like "agents compute utility of each possible action and choose the action which maximizes it", or "particles follows the trajectory which minimizes the action", and so on seem to describe a real situation, whereas they are only *substrates of modeling*. They are illustrations of the underlying classification principle which actually acts independently of how we illustrate it. From an epistemological viewpoint, the important point is not the "as-if", but the "how-if"; not whether forces exist and do act on material points, or whether agents are really computing utilities, but *how* they are assumed to do it. The answer to this how-question should highlight a classification of phenomena within a finite set of classes of models – a finite set of answers to this how-question – and this classification should not depend on the specific hypotheses used to represent it. The ontological status of forces, utilities and any not directly observable quantity is a legitimate question from a philosophical point of view. Our point is that if the criterion above is satisfied, then these models are epistemologically acceptable, and their epistemological status does not depend on the answer to the ontological question – even if the latter reflection is often fruitful.

Yet, it is true that even fundamental principles can be "explained", being derived from a more general one. We agree that it is even a really good thing from an epistemological point of view. Our point here is that, first, this derivation is not necessary to attribute a certain epistemological value to a set of general principles, a non trivial classification as described above is already sufficient. If it was not the case, we would be stuck in an infinite regression argument. Second, in the case where such a derivation of general principles $\mathcal{P}$ by more fundamental ones $\mathcal{P}'$ is possible, what is explained is no more the same set of empirical facts $\mathcal{P}$ explained, but a kind of higher-order regularity. Thus, in both cases this remark does not impinge on our core argument.

A useful analogy Maps, and more specifically metro maps, provide a useful analogy to make our point clearer.[7] A metro map can be seen in some sense as a model of the metro system it aims at representing. Some general principles give the general rules for drawing the map: a metro system is represented as a set of points tied to each other by a straight line. Some contextual hypotheses make the connection between elements on the maps and concrete elements, like: points represent metro stations and are labeled accordingly. Finally, some theoretical hypotheses give a certain form to the map: for example, a given order between stations and intersection points. From this model, then one can make some non trivial predictions: if one goes in at station S, then he can go out at station S' by taking this line, changing at this station, in this direction, and so on. However, the metro map, in order to be drawn on a readable way, also rests on some assumptions which do not correspond (and the aim of which is not to correspond) to any feature of the real metro system. These are what Mary Hesse called "negative analogies" [Hesse, 1963], properties *we know* belong to the model and not to the system modeled, in contrast with "positive analogies", properties we know belong to both, and to "neutral analogies", properties we do not know yet if they belong to the system being modeled. As she stated [Hesse, 1963, p. 9-10]: "When we consider a theory based on a model as an explanation for a set of phenomena, we are considering the positive and neutral analogies, not the negative analogy, which we already know we can discard." For instance, the form of the lines between stations are totally arbitrary, as long as the order (i.e. the structure) of stations is preserved. Besides, there often exist several different representations of a given metro system, with different forms of lines, colors or other features of this sort. Criticizing realisticness of theoretical hypotheses in models (for instance, utility maximization in rational choice models) is like criticizing the fact to represent lines between stations as having a form which is knowingly false, e.g. a straight line. In others words, some important features of a metro map (as the form of lines between stations) are not empirical statements about the real metro system but only a way of representing a certain structure, and this structure is the genuine empirical commitment of the model. Notice that the form of lines between stations could be an empirical statement, and that it has the same form as any other hypotheses in the model, that is why the confusion is easy. We do not even need to model the way metro trains actually circulate for the map to reach its epistemological function. That is why a realistic or plausible psychological theory of action for the modeling of social phenomena is an unnecessarily demanding epistemological criterion. It amounts to demand for the description of how metro trains do work in order to be able to draw a metro map for the purpose of circulating between stations. Of course, it is *easier* to draw an empirically correct metro map if we know that a metro is actually a train circulating in some precise way rather than if we think that it is a giant creature from the depths wandering underground. Likewise, it is probably easier to build up a sociological model if some psychological models of human behavior are available and robust. Our point is, again, that this is not a necessary criterion from an epistemological viewpoint.

[7] I am indebted to my colleague Antoine Brandelet for enlightening discussions about this kind of analogies and particularly about Mary Hesse's work.

Likewise metro lines have to be represented having a certain form (for instance, as straight lines), agents in rational choice models have to be represented acting in some way (for instance, maximizing a certain function over possible choices). As in physical theories, RCT can also be formulated different ways, assuming different kinds of entities and mechanisms. Let $\mathcal{A}$ be a set of possible actions on outcomes Ω with associated probabilities. If a preference relationship $\leq$ is attributed to $\mathcal{A}$ such that it satisfies some basic properties like completeness, transitivity, continuity and independence (see e.g. [Wheeler, 2020, pp. 4-5]), then it is possible to show, and this is actually a fundamental result in decision theory [Neumann and Morgenstern, 2007], that is is always possible to associate to it a function like (2) such that the preference relation $\leq$ is represented as the usual order on real numbers. Thus, expected utility theory is just a particular representation of a rationality principle. RCT can also be formulated in a more geometric way [Ryan, 2008, Durand et al., 2015]. In the latter, actions (lotteries) are vectors of a space of dimension $n = |\Omega|$, lying in the unit simplex in $\mathbb{R}^n$. The utility assigned to each outcome is represented as a vector and the expected utility of an action is the scalar product of this action and the utility vector. Again, in the light of previous reflections, we should not take these particular objects too literally, but only as possible *representations*.

These representations do not have to be realistic, or consistent with what is known about human behavior. What really counts is what types of utilities are taken into account in the optimization process, and *whether these types of utilities are already used in other similar contexts*, not whether the optimization process is realistic or not. Our central point is the following: basic assumptions (except contextual assumptions) in theoretical models do not need to be realistic because, despite their form, *they are not genuine empirical statements* but only a way of embodying a certain *classification principle* which is the actual empirical proposal to be tested. We thus disagree with epistemological positions which demand, for theoretical models to be acceptable, that their basic assumptions be "realistic" as in [Hedstrom, 2005]. Yet, we also disagree with purely instrumentalist positions as developed in [Friedman, 1953] for which empirical adequacy is a primary and central criterion. Indeed, from our viewpoint empirical success is not a sufficient condition for a theoretical model to be acceptable if this success is not related to the highlighting of a non trivial classification which does not depend on the particular formulation of the theoretical model. Concretely, in the framework of RCT, "the relevant question to ask" [Friedman, 1953, p. 15] to justify the use, for instance, of a certain form of utility is whether this form is otherwise used in a similar situation, i.e. for empirical facts or phenomena which belong to the same class.

6 Conclusion

Formal theoretical approaches like RCT are thus often trapped in a classical dilemma between empirically successful but unrealistic models (as-if models) and more detailed but trivially successful ones (just-so stories). In this paper, we proposed a solution to

this dilemma based on a *structural invariance* epistemological criterion: what counts are not general principles or theoretical hypotheses *per se*, but the structure emerging from the constraints imposed to them in order to explain a given set of empirical facts. Basic hypotheses of these models are not genuine empirical statements, despite their form, but only modeling substrates used to embody a classification principle leading to a structure which does not depend, at the end, on their particular formulation.

An important criticism which could be raised against our proposal is the fact that the social sciences have a specificity with respect to physical sciences: unlike the latter, in the social sciences we can have access to reasons why individuals do what they do. Indeed, unlike electrons or black holes, we can directly ask people or even doing introspection, as stated e.g. in [Herfeld and Ivanova, 2020, p. 4]: "Unlike in the natural sciences, we come to know principles that govern human behavior by 'immediate acquaintance' with generally available experience." This particular epistemic connection with the social sciences' objects of study indeed seems to be a strong limitation of our argument.

We see two possible answers to this criticism. Notice that they consist in whole topics of reflections in themselves, so we only sketch the general idea. First, is it really true that we have access to the reasons governing the behavior of individuals? We do have access to their *rationalization*, i.e. to the reasoning they build to explain what they do, but several pieces of evidence show that they are often distinct from the actual causality relationships at work. For example, Collier and Hoeffler's Greed or Grievance model [Collier and Hoeffler, 2004] explores the basic reasons why people engage in rebel army in the context of civil wars and compare a greed model (people engage for their opportunity costs are very low) and a grievance model (people engage for they feel political grievance against the state or regular army) and show that the relationship between these two possible reasons to engage is far from being obvious, and that greed is often the first trigger of involvement. However, asking directly people in this case, they could rationalize and focus more on grievance narratives. Social identity theory in psychology [Turner and Oakes, 1986] is another example for which people are used to rationalizing their membership to a given group even if they were assigned randomly. These discrepancies between what people tell to the researchers and the macro findings of these models suggest that although having direct testimonies and interviews is a necessary step to social inquiry, reasons that are given by people about their behavior are not necessarily actual reasons the social scientists aim at discovering.

Second, this criticism suggests that this particular epistemic connection provides more direct data than statistical or formal approaches. However, even the most direct field approach or interview is necessarily laden with theoretical assumptions, even if only basic empirical categories to think about social reality. Moreover, any observation implies an interpretation within a certain theoretical framework. In this case, it implies some assumptions about psychological features of people. However, psychology is a scientific field too, and thus some epistemological criteria apply to its models or theoretical frameworks in order to select the best one, for example their empirical success. Yet, psychological features empirically observed are never totally independent from social ones. Thus, as well as psychological assumptions would be necessary to build a sociological

theory, the latter would probably be also necessary in order to correctly interpret the findings of psychological sciences – distinguishing social from intrinsically psychological features, for example. This strong entanglement suggests that micro-features (psychological models or any "direct" observations) have no epistemic primacy on social features. Thus, from this viewpoint, social sciences theoretical framework should enjoy a certain epistemological autonomy with respect to micro-analysis, even if both approaches have necessarily to nurture each other.

Another philosophical criticism to our proposal could be that our epistemological viewpoint actually reduces to a antirealistic or instrumentalist account of social theories. However, our approach focuses on epistemological features, i.e. concerning the justification and selection of models, and not about what the empirical success of these models entails about the truth of their statements or the existence of the basic entities they postulate. As already mentioned, the word "realisticness" used in this kind of debate often simply means "consistency with what is known otherwise". However, let us notice that the epistemological criterion presented here rests on the detection of invariant epistemological structures. Thus, our epistemological criterion could perfectly be compatible with a certain form of *structural* realism [Frigg and Votsis, 2011, Ladyman, 2023]. Obviously, this goes beyond the scope of this paper and we let it for a possible future work.

References

Richard Arnowitt, Stanley Deser, and Charles W. Misner. Republication of: The dynamics of general relativity. *General Relativity and Gravitation*, 40(9):1997–2027, aug 2008. doi: 10.1007/s10714-008-0661-1.

Nicholas C Barberis. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–196, feb 2013. doi: 10.1257/jep.27.1.173.

Gary S. Becker. Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2):169–217, 1968.

Rolf Becker. Explaining educational differentials revisited: an evaluation of rigorous theoretical foundations and empirical findings. In *Handbook of Sociological Science*, pages 356–371. Edward Elgar Publishing, jun 2022. doi: 10.4337/9781789909432.00029.

Ken Binmore. *Rational Decisions*. Princeton University Press, 2011. ISBN 9780691149899.

Alisa Bokulich. Models and explanation, 2017.

R. Boudon and S.M. Lipset. *Education, Opportunity, and Social Inequality: Changing Prospects in Western Society*. Wiley, 1974.

Raymond Boudon. *Theories of social change*. University of California, 1986. ISBN 0520057597.

Raymond Boudon. The 'cognitivist model': a generalized 'rational-choice model'. *Rationality and Society*, 8(2):123–150, may 1996. doi: 10.1177/104346396008002001.

Richard Breen and John H. Goldthorpe. Explaining educational differentials. *Rationality and Society*, 9(3):275–305, aug 1997. doi: 10.1177/104346397009003002.

William A. Brock and Steven N. Durlauf. A formal model of theory choice in science. *Economic Theory*, 14(1):113–130, 1999. ISSN 09382259, 14320479.

J. S. Coleman. *Foundations of social theory*. Harvard University Press, 1990.

James Samuel Coleman and Thomas J. Farraro. *Rational Choice Theory*. Sage Publications, Inc, 1992. ISBN 9780803947610.

Paul Collier and Anke Hoeffler. Greed and grievance in civil war. *Oxford Economic Papers*, 56(4):563–595, 2004.

Stefano DellaVigna. Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–372, may 2009. doi: 10.1257/jel.47.2.315.

Franz Dietrich and Christian List. A reason-based theory of rational choice. *Noûs*, 47 (1):104–134, 2013. ISSN 00294624, 14680068. URL http://www.jstor.org/stable/43828819.

Franz Dietrich and Christian List. Reason-based choice and context-dependence: an explanatory framework. *Economics and Philosophy*, 32(2):175–229, feb 2016. doi: 10.1017/s0266267115000474.

François Durand, Benoît Kloeckner, Fabien Mathieu, and Ludovic Noirie. Geometry on the utility space. pages 189–204, November 2015. doi: 10.1007/978-3-319-23114-3_12.

Jürgen Ehlers. Survey of general relativity theory. In *Astrophysics and Space Science Library*, pages 1–125. Springer Netherlands, 1973. doi: 10.1007/978-94-010-2639-0_1.

Daniel Friedman. *Risky Curves On the Empirical Failure of Expected Utility*. Taylor & Francis Group, 2017. ISBN 9781138096462.

Milton Friedman. The methodology of positive economics. In Milton Friedman, editor, *Essays in Positive Economics*, pages 3–43. University of Chicago Press, 1953.

Roman Frigg. Models and fiction. *Synthese*, 172(2):251–268, mar 2009. doi: 10.1007/s11229-009-9505-0.

Roman Frigg and Stephan Hartmann. Models in Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition, 2020.

Roman Frigg and Ioannis Votsis. Everything you always wanted to know about structural realism but were afraid to ask. *European Journal for Philosophy of Science*, 1(2):227–276, may 2011. doi: 10.1007/s13194-011-0025-7.

Ronald N. Giere. An agent-based conception of models and scientific representation. *Synthese*, 172(2):269–281, apr 2009. doi: 10.1007/s11229-009-9506-z.

Herbert Gintis. *The bounds of reason*. Princeton University Press, 2009. ISBN 9780691140520.

John H. Goldthorpe. Rational action theory for sociology. *The British Journal of Sociology*, 49(2):167, jun 1998. doi: 10.2307/591308.

M. Göckeler and T. Schücker. *Differential Geometry, Gauge Theories, and Gravity*. Cambridge University Press, 2011. ISBN 9780511628818.

Philip A Haile, Ali Hortaçsu, and Grigory Kosenok. On the empirical content of quantal response equilibrium. *American Economic Review*, 98(1):180–200, feb 2008. doi: 10.1257/aer.98.1.180.

Michael Hechter and Satoshi Kanazawa. Sociological rational choice theory. In *Rational Choice Sociology*. Edward Elgar Publishing, dec 2019. doi: 10.4337/9781789903256.00007.

Peter Hedstrom. *Dissecting the Social*. Cambridge University Press, nov 2005. doi: 10.1017/cbo9780511488801.

Peter Hedström. Coda - the past and future of analytical sociology. In *Research Handbook on Analytical Sociology*. Edward Elgar Publishing, dec 2021. doi: 10.4337/9781789906851.00035.

Catherine Herfeld and Milena Ivanova. Introduction: first principles in science—their status and justification. *Synthese*, 198(S14):3297–3308, jul 2020. doi: 10.1007/s11229-020-02801-1.

Gudmund Hernes. We are smarter than we think. *Rationality and Society*, 4(4):421–436, oct 1992. doi: 10.1177/1043463192004004005.

Mary B. Hesse. *Models and Analogies in Science*. University of Notre Dame Press, 1963.

Kirill Krasnov. *Formulations of General Relativity Gravity, Spinors and Differential Forms*. Cambridge University Press, 2020. ISBN 9781108674652.

Clemens Kroneberg and Frank Kalter. Rational choice theory and empirical research: Methodological and theoretical contributions in europe. *Annual Review of Sociology*, 38(1):73–92, aug 2012. doi: 10.1146/annurev-soc-071811-145441.

Joost Kruis, Gunter Maris, Maarten Marsman, Maria Bolsinova, and Han L. J. van der Maas. Deviations of rational choice: an integrative explanation of the endowment and several context effects. *Scientific Reports*, 10(1), oct 2020. doi: 10.1038/s41598-020-73181-2.

James Ladyman. Structural Realism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition, 2023.

Imre Lakatos. *The Methodology of Scientific Research Programmes: Philosophical Papers*, volume 1. Cambridge University Press, 1978.

Gianluca Manzo. Educational choices and social interactions: A formal model and a computational test. In *Comparative Social Research*, pages 47–100. Emerald Group Publishing Limited, jan 2013a. doi: 10.1108/s0195-6310(2013)0000030007.

Gianluca Manzo. Is rational choice theory still a rational choice of theory? a response to opp. *Social Science Information*, 52(3):361–382, aug 2013b. doi: 10.1177/0539018413488477.

Gianluca Manzo. *Analytical Sociology. Actions and Networks.* Wiley series in computational and quantitative social science., 2014.

Gianluca Manzo. *Research Handbook in Analytical Sociology*. Edward Edgar Publishing, 2021.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, jul 1995. doi: 10.1006/game.1995.1023.

Ivan Moscati. Behavioural and heuristic models are as-if models too – and that's ok. *Economics and Philosophy*, pages 1–31, apr 2023. doi: 10.1017/s0266267123000093.

John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior (Commemorative Edition) (Princeton Classic Editions)*. Princeton University Press, 2007. ISBN 9780691130613.

Mancur Olson. *The logic of collective action*. Harvard University Press, 1971. ISBN 0674537513.

Karl-Dieter Opp. What is analytical sociology? strengths and weaknesses of a new sociological research program. *Social Science Information*, 52(3):329–360, aug 2013. doi: 10.1177/0539018413483939.

Lionel Page. *Optimally Irrational*. Cambridge University Press, nov 2022. doi: 10.1017/9781009209175.

Karl Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London, England: Routledge, 1962.

Stathis Psillos. *Scientific realism*. Routledge, 1999. ISBN 0415208181.

Matthew J. Ryan. Generalizations of SEU: a geometric tour of some non-standard models. *Oxford Economic Papers*, 61(2):327–354, aug 2008. doi: 10.1093/oep/gpn027.

H.A. Simon. *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in Society Setting*. Continuity in administrative science. Wiley, 1957. URL `https://books.google.fr/books?id=_w1gAAAAIAAJ`.

Neil J. Smelser. The rational choice perspective. *Rationality and Society*, 4(4):381–410, oct 1992. doi: 10.1177/1043463192004004003.

Daniel F. Styer, Miranda S. Balkin, Kathryn M. Becker, Matthew R. Burns, Christopher E. Dudley, Scott T. Forth, Jeremy S. Gaumer, Mark A. Kramer, David C. Oertel, Leonard H. Park, Marie T. Rinkoski, Clait T. Smith, and Timothy D. Wotherspoon. Nine formulations of quantum mechanics. *American Journal of Physics*, 70(3):288–297, mar 2002. doi: 10.1119/1.1445404.

Frederick Suppe. Theory identity. In William H. Newton-Smith, editor, *A Companion to the Philosophy of Science*, pages 525–527. Wiley-Blackwell, 2000.

Patrick Suppes. Models of data. In *Logic, Methodology and Philosophy of Science, Proceeding of the 1960 International Congress*, pages 252–261. Elsevier, 1966. doi: 10.1016/s0049-237x(09)70592-0.

Mauricio Suárez. *Fictions in science*. Routledge, 2009. ISBN 9780415990356.

John C. Turner and Penelope J. Oakes. The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology*, 25(3):237–252, sep 1986. doi: 10.1111/j.2044-8309.1986.tb00732.x.

Andreas Tutić. Revisiting the breen–goldthorpe model of educational stratification. *Rationality and Society*, 29(4):389–407, nov 2017. doi: 10.1177/1043463117734177.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, oct 1992. doi: 10.1007/bf00122574.

Gregory Wheeler. Bounded Rationality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

Per-Olof H. Wikström and Clemens Kroneberg. Analytic criminology: Mechanisms and methods in the explanation of crime and its causes. *Annual Review of Criminology*, 5 (1):179–203, jan 2022. doi: 10.1146/annurev-criminol-030920-091320.

Rasmus Grönfeldt Winther. The structure of scientific theories. 2016.

Reinhard Wippler. The structural individualistic approach in dutch sociology. towards an explanatory social science. *Netherlands (The) Journal of Sociology anc Sociologia Neerlandica Amsterdam*, 14(2):135–155, 1978.

Petri K Ylikoski. *Understanding the Coleman boat*, pages 49–63. Edward Elgar, 2021.