The Authority of Law and Indirect Theories of Decision-Making

GIUSEPPE **ROCCHÈ**
University of Palermo, Italy.
E-mail: giuseppe.rocche@unipa.it

ABSTRACT

This paper focuses on indirect theories of decision-making, i.e. those theories that tell people to use a suboptimal decision-making procedure because in the long run this will be the best way to achieve the goals established by the theory. This scheme is central in two rather distant philosophical theories. The first is Raz' "Service Conception of Authority" – centred on the idea of exclusionary reasons –, while the second is Parfit's analysis of the structure of self-interest theory and consequentialism. The comparison is helpful because it brings to light some stark differences between two kinds of indirect theories: on the one hand, indirect theories that replace counter-intuitive decision-making methods with intuitive decisionmaking methods, and in addition have the possibility of being self-effacing and esoteric (concealing suboptimality); on the other hand, indirect theories that impose a counter-intuitive decision-method and are "genealogical". It is argued that obedience to authority and rule-based decision-making are parts of an indirect theory belonging to this second kind, and this may help explain the psychological difficulty people have in understanding the ideal of Rule of Law.

Questo articolo analizza le "teorie indirette della decisione", ossia quelle teorie che prescrivono all'agente di utilizzare una procedura decisionale sub-ottimale, sul presupposto che nel lungo periodo seguire la procedura sub-ottimale sia il miglior modo di raggiungere gli obbiettivi che sono fissati dalla teoria. L'idea di teoria indiretta assume rilievo centrale in due riflessioni filosofiche lontane tra loro. La prima è la teoria dell'*autorità come servizio* di Raz – incentrata sull'idea di ragioni escludenti –, mentre la seconda è rappresentata dall'analisi della teoria dell'interesse personale e del consequenzialismo sviluppata da Parfit. Un raffronto tra queste due riflessioni è utile in quanto fa affiorare due tipi di teorie indirette: da un lato, le teorie indirette che prescrivono di sostituire un metodo di decisione contro-intuitivo con uno intuitivo, e che, inoltre, "cancellano" il carattere sub-ottimale della procedura decisionale prescritta; dall'altro lato, le teorie indiretta che prescrivono l'adozione di un metodo di decisione che è esso stesso contro-intuitivo, e non ne nascondono il carattere sub-ottimale. Si sostiene che l'obbedienza all'autorità e il modello di decisione sottoposto a regole sono parte di una teoria indiretta di questo secondo tipo, e che questa loro caratteristica può aiutare a spiegare la resistenza psicologica che le persone possono avere nel comprendere l'ideale dello Stato di Diritto e battersi per esso.

KEYWORDS

Raz, rule of law, exclusionary reasons, service conception, Parfit, consequentialism

Raz, Stato di Diritto, ragioni escludenti, autorità come servizio, Parfit, consequenzialismo

DIRITTO & QUESTIONI PUBBLICHE / CLEAR
Rule of Law: in Books, in Minds
Special Publication / February, 2026

ISSN 1825-0173

Questa Special Publication della rivista Diritto & Questioni Pubbliche è stata finanziata dall'Unione Europea – NextGenerationEU a valere sul Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Istruzione e ricerca – Componente 2 Dalla ricerca all'impresa – Investimento 1.1, Avviso Prin 2022 indetto con DD N. 104 del 2/2/2022, dal titolo "Il concetto di Stato di diritto: prospettive analitiche ed empiriche (CLEAR)", codice proposta 20225TJJSY – CUP J53D23005150006.

The Authority of Law and Indirect Theories of Decision-Making

GIUSEPPE ROCCHÈ

1. *Introduction* – 2. *Exclusionary Reasons and Indirect Theories of Decision-making* – 3. *The Motivational Account* – 4. *Theories of Practical Reason and Methods About Decision-making* – 5. *Parfit: Theories That Are Indirectly Self-defeating* – 6. *Authoritative Directives, Rules, and Indirect Models of Decision-making* – 7. *Conclusions.*

1. *Introduction*

The idea behind this essay is that sometimes practical reasoning must get worse to get better. Why is this so and when is this the case? The answer to this question has to do with the limitations of people's rationality and with different strategies to cope with these limitations. I define "indirect theories of decision-making" as those theories that set certain goals to be achieved and tell people not to be motivated by the direct pursuit of these goals, but rather to pursue something else, because that is the best way to achieve the goals. In the following sections we will focus on two philosophical traditions in which this scheme is instantiated. The first is the theory of authority in general, and in particular the theory of the authority of law. The second is the debate about the structure of consequentialism, one of the most influential moral theories. The first strand has Joseph Raz as its champion, while, regarding the latter field of enquiry, the analysis will turn to Derek Parfit's sophisticated treatment of the issue.

Indirect theories of decision-making tell people not to look at practical reasons "in a broad frame". Instead, these theories tell people to focus on some reasons neglecting others, on the premise that this sub-optimal disposition will lead, at the aggregate level or in the long run, to the best outcome. In a certain way, they exploit the scheme of "*competence without comprehension*", using Daniel Dennett's enlightening expression. Dennett's point is that many behaviours which conform to reasons are adopted without the reasons being represented in the agents' minds. Competence without comprehension is the norm in non-human life: a termite castle and Gaudí's La Sagrada Familia are very similar, but while Gaudí had in mind the reasons for designing the structure in a certain way, the reasons for the structure of the termite castle are not present in the mind of any termite[1]. Still, competence without comprehension is also at the root of many accomplishments of human societies. It led – to give a conspicuous example – to the development of the atomic bomb:

> «In 1942, Leslie Groves was appointed director of what came to be called the Manhattan Project, and in three incredibly pressured years intertwining further R&D [research and development] with the colossal (and brand new) task of refining weapons grade uranium, thousands of workers were recruited, trained, and put to work, mostly controlling the newly invented machines for separating out the isotope uranium 235, which was a fraction of 1% of the previously refined uranium 238. At the height of operations over 130,000 people worked full time on the project, *and only a tiny percentage of them had any idea at all what they were making*»[2].

[1] DENNETT 2017, 51. It must be noted that, unlike the present essay, Dennett's discourse is not focused on moral or practical reasons.

[2] DENNETT 2017, 71 (italics added). Clearly, while termites do not consciously represent their task, the workers involved in the Mahattan Project (or involved in the construction of the Sagrada Familia) had a conscious mental representation of the specific tasks they were assigned. They looked at practical reasons in a narrow frame.

The analysis of indirect theories to be presented is aimed at highlighting what may be a source of instability in the Rule of Law.

According to an empirical hypothesis, upholding the Rule of Law requires the adherence to a set of values which is cognitively more effortful than the adherence to other values. The cognitive load of Rule of Law. This work focuses on an aspect of Rule of Law which is rule-based decision-making, arguing that the adherence to this model of decision-making may be cognitively demanding, partly because rule-based decision-making is a component of an indirect theory of decision-making.

Rule-based decision-making is conceived as an improvement in practical reasoning attained through its simplification. Since it is a simplification of practical reasoning, one may wonder why the use of rules results in a cognitive load for the agents and not in a psychological relief. In a nutshell, the answer is that although the simplification is real, agents often do not feel it; rather what they do feel is that their cognitive task has become more complex. We do not necessarily feel how risky our spontaneous ways of reasoning are, and replacing a risky but spontaneous process with a less natural deliberative strategy is a hard step to take. An even more laborious step if the subject is aware of the suboptimal character of rules.

Empirically speaking, this is nothing but a hypothesis and the present paper is not meant to provide empirical support for this. The perspective is merely theoretical: the comparison between the debate on consequentialism and that about the authority of law is used to distinguish two different types of indirect theories, so as to bring out some specific problems regarding obedience to directives and rules.

The paper is structured as follows. The next three sections are devoted to the analysis of Raz' s theory of legitimate authority which offers the general framework for understanding the Rule of Law as an indirect theory of decision-making. Section 5 surveys Parfit's analysis of self-interest theory and consequentialism, contained in the first chapter of *Reasons and Persons*. Section 6 further develops Raz's account introducing the problem of rules, and then compares the two types of indirect theories emerging from the analysis.

2. *Exclusionary Reasons and Indirect Theories of Decision-making*

Does the law have or at least can it have authority over people? Raz's theory of authority is a branch of his general theory of practical reasons. His theses are not only an analytical reconstruction of the claims of law, but rather an enquiry into the moral relationship between conscientious agents and the law[3]. Because it is also a substantive enquiry and not only a piece of conceptual analysis about the claims of law, the authority of law that concerns Raz is less extensive than the authority that the law claims to have[4]. Accordingly, when Raz refers to "practical reasons" established by law, the expression is not to be understood as "legal reasons" or "reasons from the law's point of view", but as "moral reasons" or, following a more general terminology, "*real* reasons"[5].

Raz distinguishes between naked power, *de facto* (effective) authorities and legitimate authorities. The concepts of *de facto* authority and legitimate authority depend on each other. Unlike naked power (such as a criminal organization), *de facto* authority claims to be a legitimate authority, so this concept requires the concept of legitimate authority to be understood[6]. On the other hand, a likely – though not necessary[7] – condition for someone to have legitimate authority is that she is also a *de*

[3] RAZ 1986, 63, 71.
[4] RAZ 1986, 78.
[5] See GUR 2018, 10-12.
[6] RAZ 2009, 28; RAZ 2006, 1005.
[7] RAZ 2006, 1005.

facto authority[8]. The following analysis will be centred on legitimate authority, so every reference to authority without specifications must be considered a reference to legitimate authority.

According to Raz, the normative force[9] of legitimate authorities must be explained through two different types of reasons: first-order and second-order reasons for action. First-order reasons for action are reasons in favour of certain actions, while second-order reasons for action are reasons «to act for a reason or to refrain from acting for a reason»[10]. Exclusionary reasons are second-order reasons *to refrain* from acting for some reasons – that is, *to disregard* some reasons[11]. While conflicts between first-order reasons are resolved according to the strength of the conflicting reasons, conflicts between first-order reasons and second-order reasons are resolved «by a general principle of practical reasoning which determines that exclusionary reasons always prevail, when in conflict with first-order reasons»[12]. The resulting picture is that the directives issued by legitimate authorities are *pre-emptive reasons*, a conglomerate of a first-order reason for a certain action and a second-order reason to exclude some reasons – which may count against that action (or possibly in favour of it)[13] –, and the scope of exclusionary reasons is determined by those first-order reasons on which the authority has the power to pronounce[14]. This idea is expressed by the *Pre-emption Thesis*: «the fact that an authority requires performance of an action is a reason for its performance which is not to be added to all other relevant reasons when assessing what to do, but should exclude and take the place of some of them»[15].

To understand Raz's view, which can be called the "Exclusionary Model" [16] (because of the key role exclusionary reasons play in it), it is useful to contrast it with an alternative view about authoritative directives: the Weighing Model. According to the Weighing Model, authoritative directives are first-order reasons which should be added to the balance of other first-order reasons[17]. For the Exclusionary Model, instead, authoritative directives are reasons to act in a certain way and to refrain from acting on the merits of the case (that is, to act regardless of the balance of reasons[18]). The fact that a soldier has been ordered to appropriate the van of a civilian is not an additional reason in favour of the appropriation of the van that must be balanced with the reasons against, but a reason to appropriate the van and a reason not to consider the reasons within the scope of the authority in deciding what to do.

We may distinguish two very different arguments in favour of the Exclusionary Model. The first argument is conceptual, or, as it is also said, "phenomenological"[19]. If someone does what is correct on the balance of reasons by disobeying an unjust but still legitimate order it is common to respond to this conduct with a mixture of praise and blame ("decorating" and at the same time "court-martialling" the author). These mixed feelings are explained by the fact that there are reasons – like authoritative reasons – which operate at a different level than first-order reasons. According to the phenomenological argument, then, authoritative directives feature in

8 RAZ 1986, 56.
9 For the expression "normative force" I follow GUR 2018, 97 (nt.1), 100.
10 RAZ 1999 [1975, 1990], 39.
11 RAZ 1999 [1975, 1990], 38.
12 RAZ 1999 [1975, 1990], 40.
13 RAZ 1986, 42, 46; GUR 2018, 13.
14 RAZ 1989, 1194; GUR 2018, 13.
15 RAZ 1986, 46.
16 This expression has been introduced by VASSILIOU 2022, 843.
17 GUR 2018, 15. This is not the only way in which the Weighing Model may be understood. For example, Gur himself notes that for Donald Regan «authoritative directives do not themselves constitute reasons for action, but can only be indicative of existing reasons for action» (GUR 2018, 15, fn.61). It is not relevant for the present matter to complicate the issue, but I am indebted to an anonymous referee on this point.
18 RAZ 1999 [1975, 1990], 38.
19 RAZ 1989, 1165; RAZ 1999 [1975, 1990], 42 f.; see GUR 2018 (ch. V) for a general discussion of the argument and of some critical responses.

our experience as pre-emptive reasons, and not as simple first-order reasons. In other words, the Pre-emption Thesis is a conceptual or phenomenological truth.

But, while the phenomenological argument *may* prove that our concept of authority is bounded with the concept of exclusionary reason[20], it cannot establish that we actually have exclusionary reasons or pre-emptive reasons. As in moral error-theory, our beliefs, concepts and expressions may be the result of mistakes[21]. In order to prove that authoritative directives may be exclusionary reasons, a normative argument is needed: the functional argument. According to the functional argument, the function of authorities consists in making agents conform to the reasons that apply to them, and authorities can perform their function only if they are a source of pre-emptive reasons[22]. Here, the Pre-emption Thesis meets the *Normal Justification Thesis* according to which:

> «the normal way to establish that a person has authority over another person involves showing that the alleged subject is likely better to comply with reasons which apply to him (other than the alleged authoritative directives) if he accepts the directives of the alleged authority as authoritatively binding and tries to follow them, rather than by trying to follow the reasons which apply to him directly»[23].

Without a claim to completeness, political authorities provide their service because of the epistemic advantage they have over the subordinates, because they protect people from biases, because they secure schemes of social coordination and because they provide a common framework for social life: the *Service Conception of Authority*[24]. But people can harness the service provided by authority only if they treat authoritative directives as pre-emptive reasons. In fact, it is argued that the Weighing Model – the alternative theory about the normativity of law – would be incapable of guaranteeing these valuable goods in a similar way[25].

The cost of the service is that people must no longer act on their personal moral assessment of the merits of important practical issues, and must follow authoritative directives even when they are wrong. Raz is rather firm on this bitter truth: «If every time a directive is mistaken, i.e. every time it fails to reflect reason correctly, it were open to challenge as mistaken, the advantage gained by accepting the authority as a more reliable and successful guide to right reason would disappear»[26].

The emerging picture is that the Service Conception of Authority is an indirect theory of decision-making. As I said in the introduction, the general idea of the indirect theories of decision-making is that sometimes the agent's ultimate goal is better achieved if the agent, instead of directly struggling to achieve the ultimate goal, pursues another goal. This general scheme is applied to the case of authoritative directives: in order to better respond to the background first-order reasons, the agent should aim to obey the authoritative directives instead of trying to balance reasons on its own.

Nor is Raz a stranger to this logic. In Raz's words

> «Individuals should follow an indirect strategy, guiding their action by one standard in order to better conform to another»[27].

[20] By this I do not mean that the argument actually succeeds in demonstrating the phenomenological link between authority and exclusionary reasons. Moore, for example, notes that the same phenomenology of "mixed feelings" is present in conflicts between first-order reasons (MOORE 1989, 861).

[21] MACKIE 1977.

[22] GUR 2018, 97.

[23] RAZ 1986, 53.

[24] RAZ 1986, 56.

[25] RAZ 1986, 58, 75; GUR 2018, 98 ff., for the discussion of the Weighing Model see 102 ff.

[26] RAZ 1986, 61.

[27] RAZ 1986, 75.

Or, adopting the Razian distinction between conformity and compliance (which we will return to in the next section) conformity to first-order reasons – what matters – may be achieved by making people comply with authoritative directives instead of trying to comply with the balance of first-order reasons:

> «advantages [in having authorities], I will argue, are always the result of indirect strategies of conformity with reasons, i.e. maximizing conformity with reasons not by trying to comply with them, but by following someone else's judgment about what one should do (...) The two basic arguments for authority depend on its ability, through concentrating expertise on various issues, to overcome common ignorance and on its ability to help solve common difficulties in securing coordination. Overcoming both problems requires adopting an indirect approach to conformity to reasons, that is, it requires securing conformity not through an attempt to comply»[28].

Although not alien to Raz's thought, the idea of an indirect method of decision-making has not been fully developed in the context of the theory of authority. It may be useful then to draw a comparison with other debates in which the scheme of indirect methods of decision-making emerges. In this article, I will focus on Parfit's analysis of the structure of self-interest theory and consequentialism.

But before moving in this direction two additional issues must be addressed. The former has to do with the current interpretation of the exclusionary reasons, that is, the *Motivational Account*[29]. Focusing on the Motivational Account (section 3) is important in the present context to cast light on the role exclusionary reasons play in the agent's deliberative process. Still, the Motivational Account clarifies this role only to a limited extent. Therefore, section 4 focuses on certain ambiguities that the Motivational Account does not solve.

3. *The Motivational Account*

As delineated above, exclusionary reasons are reasons to disregard other reasons, that is, for not acting for certain reasons, or to refrain from acting for them. But what is the meaning of these and similar expressions? What does it mean to "disregard reasons" or "not act for a reason"?[30] These expressions need to be clarified. Raz explains his view in *Facing Up* and in *The Postscript* to *Practical Reasons and Norms*. In a few words, according to Raz "acting for a reason" means "being motivated by a reason", and then exclusionary reasons are «reasons for not being motivated in one's action by certain (valid) considerations»[31].

In order to understand the Motivational Account, we must first introduce some closely related distinctions. The first is the distinction between two subjective relations that the agent may entertain with reasons for action: conformity and compliance. The second regards the nature of practical reasons: the distinction between normative reasons and motivating reasons.

If Derek has a reason to stay home because Jane needs moral support, and Derek stays home, Derek *conforms* with that reason. In general, «people conform with a certain reason if they perform that act in the circumstance in which that reason is a reason for its performance»[32]. If Derek not only stays home, but does so in order to give Jane moral support, we say that Derek *complies* with

[28] RAZ 1999 [1975, 1990], 195.
[29] For this expression see ADAMS 2021, 236.
[30] I follow here Muffato's helpful and thorough analysis (MUFFATO 2022, 513 ff.).
[31] RAZ 1999 [1975, 1990], 185.
[32] RAZ 1999 [1975, 1990], 179.

the reason. According to Raz, reasons for action are reasons for conformity, not for compliance[33]. To be clear, they are normally also reasons for compliance, as far as compliance is a step toward conformity[34], and additionally Raz does not deny that compliance with reason can be a way to express the correct moral sensibility. Still, there are cases in which acting in a sort of unreflexive way is not only enough to do the right thing, but it is also the most desirable way to behave[35], and this may be taken as a proof (among others) that reasons for action are reasons for conformity.

Related to the same point is the distinction between normative and motivating reasons. A normative reason is a consideration counting in favour of a certain action, while a motivating reason is a reason for which, or on the basis of which a person performs a certain action[36]. The connection between the idea of motivating reason and the above distinction between conformity and compliance is that when a person is complying with a certain reason, that reason is a motivating reason for the subject[37].

Against this background, the core idea of the Motivational Account is that exclusionary reasons are reasons for not being motivated by excluded reasons, that is, for not complying with the excluded reasons, *but* are not reasons for not conforming to the excluded reasons[38]. They defeat certain considerations as motivating reasons, but not as normative reasons. Let's turn back to the example of the authoritative directive, in which the soldier has been ordered to appropriate a van belonging to a civilian. Postulating that the order is a reason for taking the van, that excludes reasons related to the military uselessness of this operation, we would say that the order defeats considerations related to military uselessness as motivating reasons. But it does not defeat them considered as normative reasons[39]. Even after the authoritative directive has been issued, military uselessness remains a normative reason not to take possession of the civilian's van.

The contrast between motivating reasons and normative reasons explains the difference between exclusionary reasons and cancelling conditions[40], also called "disabling conditions"[41]. An agent promises to do a certain action to his friend, but then the friend releases the agent from the promise. While the promise was a reason for a certain action, the release from the promise is a reason for nothing at all, but it cancels the reason generated by the promise. In this case we must neither comply with *nor* conform to the promise, since the reason has been cancelled. While the cancelling conditions defeat certain reasons *qua* normative reasons, exclusionary reasons defeat them just as motivating reasons, leaving the normative force of the excluded first-order reason untouched. In Andrew Jordan's words:

> «To sum up, the distinction between first-order reasons, exclusionary reasons, and disabling conditions could be understood as follows: First-order reasons stand in relation to actions and either favor or disfavor them. Exclusionary reasons, in contrast, stand in relation to the motives of an agent. They demand of an agent that she not act for certain first-order reasons. And disabling conditions stand in relation to first order reasons to turn off their normative force—that is, to make them non-reasons»[42].

[33] RAZ 1999 [1975, 1990], 180

[34] RAZ 1999 [1975, 1990], 182.

[35] Raz here challenges the Kantian idea according to which «only respect for the moral law is an appropriate moral motive» (RAZ 1999 [1975, 1990], 181). The passage seems to challenge a peculiar motive rather the relevance of motive in general. Be that as it may, the critique to the Kantian approach resembles Bernard Williams' thinking (see WILLIAMS 1981 [1975]; see also SMITH 1994, 75 f.)

[36] WHITING 2015, 398.

[37] MUFFATO 2022, 514 f.

[38] RAZ 1999 [1975, 1990], 194; MOORE 1989, 856 ff.; RAZ 1989, 1156 ff.

[39] ADAMS 2021, 238.

[40] RAZ 1999 [1975, 1990], 27.

[41] JORDAN 2018, 360.

[42] JORDAN 2018, 349.

The Motivational Account clarifies, from a theoretical point of view, the mixed feelings aroused by the act of disobedience that has led to the performance of the correct action on the balance of reasons. Since exclusionary reasons defeat first-order reasons only as motivating reasons and not also as normative reasons, there is both a clear sense in which the agent did what she ought to do and another clear sense in which she failed to perform her duty.

A crucial case for understanding the Motivational Account – although somewhat mysterious and convoluted – is the case of the lucky error. Again, we must imagine an authoritative directive commanding the execution of an action that is different from that pointed out by the balance of reasons. The balance of all first-order reasons and the balance of unexcluded reasons point in different directions. The directive is wrong on the merit. But, even if wrong, the directive comes from a legitimate authority and is an exclusionary reason. In a situation like this, there is no chance that the agent can conform both to the balance of all first-order reasons and to exclusionary reason, as far as the agent "reasons correctly". But assume that the agent, although not motivated by the balance of all the first-order reasons, *miscalculates* and performs the real good action, the action pointed out by the balance of the background first-order reason. In this case – Raz notes – the error is fortunate, because the agent managed to conform both to first-order reasons and to exclusionary reasons. This is because exclusionary reasons are not reasons for not conforming to the excluded reasons (in that case the situation described would have been truly tragic; there would have been no room for any lucky mistake), but merely for not complying with them, and in this case the agent has not complied with them, since she has not been motivated by the balance of all first-order reasons[43].

This is the Motivational Account. Exclusionary reasons seem to be part of a theory of decision-making, according to which authoritative directives are reasons to deliberate in a certain way, i.e. to avoid being motivated by certain considerations. Still, in the next section, it will be shown that the issue is more complicated. The Motivational Account may have left open many important questions.

4. *Theories of Practical Reason and Methods About Decision-making*

Criticizing Noam Gur's account of the normativity of law, Andrej Vassiliou appeals to a distinction between theories of practical *reason* and methods of decision-making or theories of practical *reasoning*. Vassiliou writes:

> «A theory of practical reason addresses the questions of what kinds of practical reasons obtain, when these reasons conflict with each other, how these conflicts are resolved and, most importantly, what agents ought to do in light of their different and commonly conflicting reasons. Conversely, the question that decision-making methods answer is not what agents ought to do, but how they should reach the right decision on what they ought to do»[44].

We are not interested here in the critique of Gur's model. Suffice it to say that, according to Vassiliou, the Weighing Model and the Exclusionary Model are theories of practical reason, but this fact does not settle the issue of which method of decision-making (if any) is prescribed by those theories. It may be the case that the adoption of a theory of reason favours an isomorphic method of decision-making, but Vassiliou notes – echoing here Ruth Chang's account – that this cannot be taken for granted; it is possible to advocate a certain account about reasons

[43] RAZ 1999 [1975, 1990], 185 f.
[44] VASSILIOU 2022, 946.

without a commitment in favour of the isomorphic method of decision-making[45]. For example, it is possible to adhere to the Weighing Model or in general to a model prescribing «that what we ought to do is ultimately determined by the balance of our reasons when these reasons conflict»[46], without accepting that the appropriate decision-making procedure consists in balancing the weights of the conflicting considerations – that is, without adopting the isomorphic decision-making procedure.

Following this suggestion, theories of practical reason may legitimize different methods of decision-making, even methods of decision-making which do not resemble the content of the theory of reason. In other words, methods of decision-making may be conceived as tools to fulfil the goals identified by the theory of practical reason. In this sense, when there is no isomorphism between the method of decision-making and the theses about reasons for action, the background theory of reason would be an indirect theory of decision-making.

The distinction between theories of reason and models of decision-making may be misleading[47], still Vassiliou's idea that the Exclusionary Model is a theory of reason and not a method of decision-making has a strong implication for the present analysis.

A possible way – certainly not Vassilou's – to use the distinction between theories of reason and methods of decision-making would be to say that the Service Conception of Authority is an indirect theory of decision-making consisting of two levels.

> First level (supreme level): a substantive theory of justice establishing the goals to be pursued, i.e. a theory about the background, first-order, reasons.
>
> Second level (derived level): the Exclusionary Model (according to the Motivational Account).

The first (supreme) level is that of the theory of reason, while the second (derived) level is that of the decision-making method. The Service Conception would be an indirect theory of decision-making in which exclusionary reasons feature at the derived level of decision-making, the Exclusionary Model would be a model prescribing how to deliberate. It would tell people to perform an "exclusion task", that is, to focus – with their mental control – on keeping some considerations – which for them are reasons – out of their deliberation.

Alternatively – and this may be Vassiliou's proposal – the Exclusionary Model may be understood not as a model prescribing a way to deliberate, but as a model which individuates some practical goals and legitimates the pursuit of these goals through different decision-making procedures. This means, in my opinion, that the levels of the Service Conception of Authority would not be two but three:

> First level (supreme level): a substantive theory of justice establishing the goals to be pursued, i.e. a theory about the background, first-order, reasons.
>
> Second level (derived intermediate level): the Exclusionary Model (according to the Motivational Account).

[45] VASSILIOU 2022, 954; CHANG 2016, 215 f.

[46] VASSILIOU 2022, 954.

[47] The distinction between theories of reason and models of decision-making may be problematic because it seems to imply that the methods of decision-making do not regard practical reasons at all. But Vassiliou's idea is rather that, although methods of decision-making do not give agents reasons for *action*, they give them reasons «to structure their deliberative process or the environment within which they deliberate in a particular way» (VASSILIOU 2022, 946). So understood the distinction may remain misleading as far as it seems to suggest that reasons to structure our deliberative process are not practical reasons in the full sense. It may be wondered how it would be possible to understand exclusionary reasons as reasons for action, once the Motivational Account has been adopted.

> Third level (derived last level): the set of decision-making procedures that prevent the agent from being motivated by the excluded reasons.

The Motivational Account, in itself, does not rule out this possibility. This is the ambiguity I have anticipated above. In fact, if the exclusionary reasons are reasons for not being motivated by certain reasons, this does *not* imply that the agent must be motivated by exclusionary reasons; it does not imply that the agent must perform a certain mental activity when deliberating. In other words, although exclusionary reasons do not cancel first-order normative reasons, but merely deprive first-order reasons of the property of being motivating reasons, this fact does not imply that, according to the Exclusionary Model, exclusionary reasons must be in their turn motivating reasons[48]: they may be just goals that legitimize the adoption of whatever method of decision-making may realize a state of affairs in which the agent is not motivated by the excluded first-order reasons.

The Motivational Account is ambiguous also in a second way. Even if exclusionary reasons are motivating reasons and so they must be present in the agent's deliberation, a question may be raised about which motivation is satisfactory. Is it enough to try to exclude certain considerations with the old-fashioned mental control, or must the agent prepare herself in a certain way, developing attitudes that will make her obey correctly and efficiently?[49]

So, although one might get the impression that the Motivational Account outlines a method of deliberation that people must follow when confronted with authoritative directives, the truth, in my opinion, is more complicated. The Motivational Account clarifies only that exclusionary reasons are not cancelling conditions, but it remains unclear about which role has been assigned to exclusionary reasons in the deliberative process.

The discussion of the role of exclusionary reasons in the deliberative process is pertinent for an encompassing analysis of the Exclusionary Model and the indirect methods of decision-making. The idea would be, in effect, that not only the balance of background first-order reasons, but also exclusionary reasons, can be indirectly pursued. Moreover, if the levels of the Service Conception would be three and not two, the problem would arise as to how to justify the intermediate level – the one in which the exclusionary reasons operate. If, in order to respond to reasons, we must ultimately adopt certain decision-making strategies, why should we choose them on the basis of their suitability to achieve the balance of non-excluded reasons, rather than – bypassing the intermediate level – on the basis of their suitability to make us respond to the balance of background reasons?

Still, we do not need to complicate the issue now. We may treat the Exclusionary Model as a model of decision-making, since, even if its best reconstruction would be another, it is safe to assume that this model dictates *also* – at one point or in many cases – a certain method of decision-making, whereby the agent must try to disregard through her mental control those considerations that seem to her reasons for action, but which she also believes to fall within the scope of the authoritative directive.

5. *Parfit: Theories That Are Indirectly Self-defeating*

The debate about indirect consequentialism offers a sophisticated framework to understand the more general idea of indirect theories of decision-making. Indirect consequentialism is often

[48] I thank Nicola Muffato for the enlightening conversation on this point. On the evolution and oscillations of Raz's account see MUFFATO 2022.

[49] We may distinguish here between reasons working "on the spot" and reasons for preliminary activities. I owe this suggestion to Noam Gur.

conceived as a strategy to rescue consequentialism from the problem of alienation. The premise is that direct, simple, consequentialism would be flawed because it would require moral agents to be motivated by the maximization of the good, and thus it would lead people to alienate from the things that are important to them[50]. On this premise, following Sidgwick, consequentialists reply that a «consequentialist agent need be committed to maximization of the good only as an objective criterion of rightness by which his actions can be assessed, rather than as directly providing a motive or a purpose which such an agent is to consciously adopt in performing any action»[51].

In other words, consequentialism would be an indirect moral theory, because the maximization of the good serves as the criterion of rightness for evaluating both which actions ought to be performed and which motives agents ought to possess in order to act rightly, yet at the same time, it is not necessary for the maximization of the good to figure in the agents' motivations. The maximization of the good would be the criterion which legitimizes certain actions and certain motivations but not also the motivation required. According to consequentialism agents are not required to be motivationally guided by consequentialism, and, in some cases, they are even required not to be.

In what follows, I will refer to Derek Parfit's analysis contained in the first chapter of his masterpiece, *Reasons and Persons*. Although Parfit does not use the expression "indirect consequentialism", his analysis of the problem excels in terms of thoroughness and generality. In Parfit's analysis, on one hand, alienation is not the only problem for direct consequentialism that may motivate the turn towards indirect-consequentialism, on the other, the metamorphosis in indirect theories of reasoning is not a phenomenon circumscribed to consequentialism, but also regards other theories of practical reasons, in particular self-interest theory[52].

Parfit opens *Reasons and Persons* by observing that critiques of various theories of rationality or morality rely on certain assumptions. But there is an exception. One peculiar critical argument is that the theory of rationality or morality is *self-defeating*. «This argument, uniquely, needs no assumptions. It claims that a theory fails even in its own terms, and thus condemns itself»[53].

Parfit believes that both self-interest theory and consequentialism are *indirectly* self-defeating. A theory is «*indirectly self-defeating* when it is true that, if someone tries to achieve his T-given aims [the aims given by the theory], these aims will be, on the whole, worse achieved»[54]. In terms of motivations, a theory is indirectly self-defeating when an agent who is motivated by that theory ends up achieving the aims it prescribes less effectively than she would have by being motivated by a different theory. However – this is Parfit's contention –, being indirectly self-defeating is not, as it might seem, a problem for these theories.

The analysis starts discussing the case of self-interest theory. According to self-interest theory, our aim is that our life goes as well as possible[55]. Different theories give different answers to the question about what self-interest consists of. Parfit enumerates three famous accounts: hedonism, the desire-fulfilment theory, and the objective list theory[56]. The general idea is that if the agent acts motivated by the goal that her life should go as well as possible, the goal that her life will go as well as possible will be achieved worse[57].

Parfit distinguishes in this respect between two possible sources of this failure. The first

[50] WILLIAMS 1973, 128.

[51] COCKING, OAKLEY 1995, 87. This reply is found for instance in RAILTON 1988.

[52] Cocking and Oakley survey the most famous contributions about indirect consequentialism (COCKING, OAKLEY 1995, 87). See also WILAND 2007; and MCNAUGHTON, RAWLING 2025 which though not explicitly quoting Parfit is clearly referring to his work.

[53] PARFIT 1984, 3.

[54] PARFIT 1984, 5.

[55] PARFIT 1984, 3.

[56] See also GRIFFIN 1986.

[57] PARFIT 1984, 5. At p. 8 Parfit starts talking about "motives".

possibility relates to the case in which the agent tries to follow the theory but fails, doing what is worse for her. We may call this eventuality the "performance error" problem. In the second case, by contrast, although the agent reliably avoids choices that are contrary to her own good, her life is nevertheless worse than it would have been had she been guided by dispositions other than pure self-interest. For this unwelcome result to occur, it is not necessary for the agent to be always motivated by her self-interest; rather it suffices that she is "never self-denying" – i.e., that she never does what she believes would be worse for her[58]. We may call this case, which stems from the fact that the agent is never self-denying and which does not involve any performance error, "the pure motivational case".

The case of the performance error is not so much a reason to criticise the theory as it is a reason to criticise the person who made the error. More interesting for Parfit is the pure motivational case – the case where the agent does nothing that is worse for herself, but, since she is never self-denying, she achieves her aims in terms of self-interest less than if he had different motivations.

Still, it is important to dwell on the supposedly uninteresting case of performance error. Although Parfit observes that the agent's incompetence in following the theory is not a fault of the theory but an agent's fault, it is worth noting that failing to follow self-interest theory because it is too complicated might have been, even according to Parfit, an objection to self-interest theory. The difficulty to follow a theory is after all an argument against that theory. The reason why this is not an actual concern for self-interest theory is just that, as a matter of fact, self-interest theory is not difficult to follow[59]. This detail is important because one of the limits of a decision-making process based on the balance of background first-order reasons lies in its inherent complexity. That said, though the performance error problem is nearer to the Razian analysis, in order to follow Parfit's analysis we will focus on the pure motivational case.

How could our motivations doom us in this way? An example[60] provided by Parfit to illustrate this scheme has to do with a case of rescue. Imagine that the agent is stranded in the desert and manages to stop someone who may rescue her. The agent may offer the stranger a great reward for his help. Assume next both that the agent is never self-denying and that she is "transparent", meaning that she is unable to deceive others. If she is never self-denying, it is true that later, when she will be safe at home, she will not give to the rescuer the reward. But if she knows it and she is transparent then it is also true that she is unable to promise the reward convincingly. The result is that the agent is left stranded in the desert, an outcome that is worse for her. Because of her transparency it would have been better for the agent in self-interest terms if she had been self-denying, because only in this way she could have been trustworthy[61].

Self-interest theory is then self-defeating. Anyway, as already mentioned, according to Parfit, self-interest theory is self-defeating in a way that does not falsify it. Though being never self-denying is the motivation which somehow mirrors the imperative of self-interest theory and it is in this sense a rational motivation (assuming the correctness of self-interest theory), self-interest theory does not tell people to be never self-denying; rather, it advises cultivating a set of motivations that are optimal from the perspective of self-interest. People are not told to be never self-denying, and are encouraged to be so when it serves their self-interest. Therefore, self-interest theory is not failing in its own terms[62]. The pure motivational case shows rather that there are cases in which it is rational to cause ourselves to act on a set of motives which are in themselves irrational: rational irrationality.

[58] PARFIT 1984, 6.
[59] PARFIT 1984, 5.
[60] A fanciful example, like many, but one that should make the problem clear.
[61] PARFIT 1984, 7.
[62] PARFIT 1984, 8-12.

According to the scheme of rational irrationality, the agent has self-interest reasons to have motivations leading to actions which will be irrational, but whose possession is a way to generally improve the conformity to self-interest aims: in short, the actions may be deemed irrational, but not the agent[63]. Imagine a writer whose strongest desire is that her novel be as good as possible. This agent is disposed to sacrifice her self-interest for this goal. Still, it would be worse for the writer, in self-interest terms, if her desire were weaker. This writer possesses the motivation that would be irrational for her to lose. Yet, when she works hard to the point of exhaustion and depression, it remains true that she is acting irrationally according to self-interest theory. Given self-interest theory, avoiding exhaustion and depression remains a reason for her, even though she has reasons to maintain the motivations that will lead her into those states[64]. The parallel with the Motivational Account of the Exclusionary Model is straightforward: in both cases, certain factors cease to be motivating reasons while remaining normative reasons for action, and in both cases, this is the price the agent must pay for the general conformity to reasons.

But here there is another twist. Never being self-denying might be the result of people's *belief* in self-interest theory. If we suppose that people's motivation to be never self-denying derives from their belief in self-interest theory, and that they can never change their motivation as far as they believe in self-interest theory – i.e. they cannot (intentionally) do what they believe to be irrational –, this means that the very belief in self-interest theory has a bad effect in self-interest terms, and thus self-interest theory fails in its own terms[65]. Parfit's answer to this objection is that, in this case, self-interest theory not only would give reasons to act on a set of motives incoherent with self-interest, but also would give practical reasons to believe in other theories about rationality. Self-interest theory would be *self-effacing*, a theory which removes itself from the scene, telling anyone to believe another theory, so far as having this belief is useful to have the set of dispositions that are best in self-interest terms[66]. It must be noted that this legitimization of other theories about how to live our lives does not count as a full theoretical justification of these theories, rather the lesson to learn is that self-interest theory may dictate to believe in theories which are false; self-interest theory may dictate to have false beliefs[67].

As Parfit concludes:

> «to be self-effacing is not to be self-defeating. It is not the aim of a theory to be believed. If we personify theories, and pretend that they have aims, the aim of a theory is not to be believed, but to be true, or to be the best theory. That a theory is self-effacing does not show that it is not the best theory. (...) Though S [self-interest theory] would not be failing in its own terms, it might be claimed that an acceptable theory cannot be self-effacing. I deny this claim. It may seem plausible for what, when examined, is a bad reason. It would be natural to *want* the best theory about rationality not to be self-effacing. If the best theory was self-effacing, telling us to believe some other theory, the truth about rationality would be depressingly convoluted. It is natural to hope that the truth is simpler: that the best theory would tell us to believe itself. But can this be more than a hope? Can we assume that the truth *must* be simpler? We cannot»[68].

[63] PARFIT 1984, 12-17. To be clear, acts by which one does not satisfy an immediate preference in order to be much better off in the future (such as undergoing a painful operation) are not irrationally rational acts, but rational acts in the most canonical sense. In fact, refraining from satisfying an immediate preference in order to gain a greater advantage in the future is not self-denying. An agent is self-denying when she does what she believes would be worse for her.

[64] PARFIT 1984, 14.

[65] PARFIT 1984, 11, 17 ss..

[66] PARFIT 1984, 23.

[67] Therefore, we cannot legitimize morality through self-interest, thereby reconciling moral and self-interest reasons and solving "the profoundest problem of Ethics", in Sidgwick's words (PARFIT 1984, 19).

[68] PARFIT 1984, 23.

Similar considerations apply to consequentialism[69]. The fundamental idea of consequentialism is that there is an ultimate moral aim that consists in the realization of the best possible outcome. Different versions of consequentialism offer different answers to the question about what makes an outcome good, better and best. According to hedonistic utilitarianism, for example, the best outcome is the one in which there is the greatest net sum of pleasure minus pain. But there are more sophisticated and less famous versions of consequentialism. Unlike hedonistic utilitarianism there may be versions of consequentialism which adopt a pluralist axiology; further, unlike simpler versions that rank states of affairs, there may be versions that evaluate histories of the world.
Although it may seem that consequentialism is so general as to be void, according to Parfit all forms of consequentialism share a peculiar common structure, which is condensed in the idea of *agent neutrality*[70]. A moral theory is agent-neutral if it gives a common moral aim to all agents. Assume that the moral aim which has priority over all the others is the respect for human rights. A consequentialist approach to human rights would imply that each agent ought to minimize the occasions on which human rights are violated. This means that there may be cases in which an agent must actively violate human rights to prevent a bigger violation of human rights perpetrated by someone else. Agent neutrality is contrasted with *agent relativity*, which is then the hallmark of the non-consequentialist theories, also known as "deontological theories". A theory is agent-relative if it gives different people different aims. Considering human rights once again, according to an agent-relative theory of rights the agent must refrain *herself* from violating rights, and she is not permitted to violate rights even in order to prevent greater violations committed by others. Nozick's account of fundamental rights according to which rights are side constraints to actions is the typical example of this type of theory[71].

Having clarified the notion of consequentialism, Parfit asks whether consequentialism is indirectly *collectively* self-defeating. A theory is indirectly collectively self-defeating «when it is true that, if several people try to achieve their T-given aims, these aims would be worse achieved»[72]. As in the case of self-interest theory, the concern is that if agents were motivated (disposed) to bring about the best outcome, the resulting outcome would not be the best achievable, even if they successfully performed the correct actions (even in the absence of performance errors). The failure might come not from the agents' actions, but rather from their motivations. The difference with the case of self-interest theory is that the failure of the aim given by the theory would occur only if many people were motivated to bring about the best outcome.

Why should a collective motivation to bring about the best outcome make the outcome worse? The idea is that agent-neutral consistent and enduring motivations would result in the suppression of the agent's desires. And if all or most of the agents suppress their desires and life-plans, this will result in a massive contraction of the aggregated sum of happiness[73].

Although this is the central case Parfit is going to discuss, it must be noted – symmetrically with what has been said regarding self-interest theory – that consequentialism may be self-defeating also because of the problem of performance errors. One example given by Parfit concerns the likelihood that a person may deceive herself about the outcomes of her actions. «If we want someone to be dead, it is easy to believe, falsely, that this would make the outcome better. It would therefore make the outcome better if we were strongly disposed not to kill, even when we believe that this would make the outcome better»[74]. Once more, a case which very closely resembles the type of concerns which supported the adoption of the Exclusionary Model.

[69] PARFIT 1984, 24 ff.
[70] See for example PORTMORE 2001.
[71] NOZICK 1975.
[72] PARFIT 1984, 27.
[73] We found this critique in WILLIAMS 1973; for an analysis of the problem see ADAMS 1976.
[74] PARFIT 1984, 28.

As in the case of self-interest theory so in the case of consequentialism, the fact that consequentialism *may be* indirectly self-defeating does not prove that the theory fails in its own terms. Unlike in the case of self-interest theory, Parfit believes that in the real world consequentialism is not indirectly self-defeating. For, to be self-defeating most people should be disposed to realize the best possible outcomes; but the reality is that very few people are disposed to produce the best outcome in agent-neutral terms. But, more generally, like self-interest theory, consequentialism does not require that people are motivated by the aim that outcomes be as good as possible, but rather that people have the best set of motives in consequentialist terms, that is the set of motives which will make outcomes as good as possible[75]. This means that consequentialism requires people to cultivate a set of motivations that, although they may occasionally lead to wrong actions, will, on the whole, result in the best possible outcome. The acquisition of these motivations is good overall and then the agents who acquire them are blameless when they do the wrong thing: blameless wrongdoing.

A useful example is that of a mother who can either benefit her child or a stranger, where the benefit that she could give to the stranger is bigger than that she could give to her child. When the mother benefits her child, she is acting wrongly in consequentialist terms. However, since it would have been wrong for her – again in consequentialist terms – to lose her love for her child, the wrongness of her action is the product of what is overall good. The action remains wrong, but the agent should not be considered immoral for having performed it[76]. Once again, the situation described is very similar to the Motivational Account of the Exclusionary Model: what was a normative reason for a certain action remains a normative reason – since the ultimate criterion of rightness has not been changed –, but the agent ought not to be motivated by that reason, and, again, this is the price the agent must pay for the general conformity to reasons.

Because consequentialism may tell people not to be motivated to maximize the good, and it may be the case that people are unable to change their motivations in the requested way as long as they believe in consequentialism, consequentialism may tell people to believe in another theory. For example, may tell people to believe in an improved version of common-sense morality[77]. In this case, consequentialism would be self-effacing. Still, Parfit believes that it is not the case that consequentialism is wholly self-effacing, rather it is more plausible that consequentialism is partially self-effacing and partially *esoteric*: it is morally good that consequentialism be adopted by an enlightened minority who do not disclose the theory to the vast, ignorant majority. While the image of this Government House Utilitarianism – to use Williams' caustic expression[78] – is disconcerting for many people, once again Parfit believes that depressing truths in ethics are still truths, and seems to side with Sidgwick, who regretted the esoteric character of consequentialism, «but he did not think regret a ground for doubt»[79].

To conclude, the charge according to which self-interest theory and consequentialism fail in their own terms has been rebutted by Parfit. In doing so, Parfit has shown that self-interest theory and consequentialism have a tiered structure, and this is why I think it is meaningful to speak of "indirect" theories of decision-making. They are indirect theories of decision-making because they do not tell people to apply their fundamental values in the decision-making processes, but they legitimize other decision-making models (some motivations) as the best way to reach the best outcome. Our reasoning must get worse to get better in the long run.

[75] Parfit 1984, 28.
[76] Parfit 1984, 31-40.
[77] Parfit 1984, 40.
[78] Williams 1985, 108; see also Sen, Williams 1982, 16.
[79] Parfit 1984, 41.

Using the distinction between theories of practical reason and theories of reasoning, self-interest theory and consequentialism must be understood as theories of practical reason – that is, as criteria for the rightness of action – which, in certain cases, instruct individuals not to adopt them as their own theory of reasoning. When it comes to self-interest theory and consequentialism Parfit has denied that there is an isomorphic relation between the criteria of correctness and the decision-making models.

6. *Authoritative Directives, Rules, and Indirect Models of Decision-making*

Although one of the biggest problems of the modern world is that obedience is all too easy[80], it may also be true that the authority of law and in particular the ideal of Rule of Law are cognitively burdensome. Parfit's thoughtful analysis of self-interest theory and consequentialism provides a framework to understand the cognitive difficulties related to obedience in general, and obedience to rules in particular. This conclusive section compares indirect consequentialism and the Service Conception which adopts the Exclusionary Model, dwelling on the problem of rules[81].

Both in indirect consequentialism and in the Service Conception the decision-making process must in some sense get worse in order to get better. In both cases the decision-making process worsens from the agent's point of view because:

(1) The agent has reasons not to be motivated by the criteria of correctness for action directly – in Raz's terms, the agent has reason not to be motivated by some reasons, denying the seemingly ubiquitous principle of practical reasoning according to which «one ought, all things considered, to do whatever one ought to do on the balance of reasons»[82]; and this – as a claim about decision-making – means that the agent knows that she has reasons not to be motivated by her *personal moral assessment* of the background reasons.
(2) The agent knows that, if she acts on the right motivations, there will be cases in which she will do the wrong thing, while she would have done the right thing had she acted directly on the criteria of correctness – that is, on the basis of her personal moral assessment of the background reasons.

In the case of consequentialism, (2) occurs when the agent, justified in following a refined version of common-sense morality, performs an action that is right by common-sense standards but not according to consequentialism. There are two different situations that make (2) true when the Service Conception tells us to adopt the Exclusionary Model of decision-making.

(2.1) Although the command may come from a legitimate authority which is operating within the scope of its authority, the command may be wrong (even terribly wrong). In this case, though the balance of reasons points towards action x, the agent has a first-order reason to do non-x and an exclusionary reason not to be motivated by those reasons for x over which the authority has the power to pronounce.
(2.2) In the second situation the authoritative directive is not generally and abstractly wrong, but is a *rule* that in the specific case misfires. The legal rule is not morally objectionable in itself. However, certain circumstances arise in the particular case such that the agent would be doing the wrong thing (even something terribly wrong) by obeying the rule[83].

[80] See for example SMEULERS 2019.
[81] I will set aside self-interest theory.
[82] RAZ 1999 [1975, 1990], 36.
[83] These two cases are adaptations of Gur's Situation 1 and Situation 2 (GUR 2018, 22 f.).

The problem of rules (2.2) within the theory of authority deserves attention.

Although not all authoritative directives are rules, as the concept also encompasses particular norms[84], many of them are. And while the law does not necessarily consist of rules imposing substantive constraints[85], these rules are crucial in many legal systems. A necessary condition for the Rule of Law – though by no means sufficient – is the rule *by* law, that is the use of orders that are general, clear, prospective, public, and relatively stable[86], imposing substantive constraints. In other words, the government of people through legal rules.

What is peculiar of rules is that they have the capacity to generate wrong results even if their creator never intended to produce them[87]. This is so because of the way in which rules are (ideally) manufactured.

Following Schauer, a rule is created through a process of generalization from particular cases, where the resulting general factual predicate of the rule – the antecedent of the conditional statement, or *protasis*– is chosen according to the purpose, the rationale, the "justification" of the rule «the evil sought to be eradicated or the goal sought to be served»[88]. A black dog enters the restaurant barking and disturbing customers. Generalizing from this particular case, if our purpose is that of satisfying customers protecting them from various annoyances, it is reasonable to abide with a rule according to which "dogs are not allowed", instead of the rule "black things are not allowed", since dogs, and not black things, are often cause of annoyance for the customers of restaurants.

But while the general factual predicate must be related to the production of the feared or desired consequence, it «is not a statement of the individually necessary and jointly sufficient conditions» for that, as it merely describes a set of facts that bear a relation of "probabilistic causation" to the justification [89]. While "Dogness" is probabilistically related to annoyances, it is neither a necessary nor a sufficient condition for the occurrence of annoying events in the restaurant. Because rules are the result of probabilistic generalizations, they are over-inclusive with respect to their justification, in the sense that they encompass cases that should not be included according to the justification of the rule – for instance, well-behaved dogs, blind dogs or the Queen's corgis. Moreover, the mere fact that rules are generalizations from specific cases – regardless, this time, whether they are probabilistic generalizations or not – makes it the case that they are also under-inclusive, failing to encompass cases which should be included according to the rationale – like a pet bear or a giant snake[90].

This means that rules sometimes do not serve their purpose and may well impede it actively. Rule-based decision-making is necessarily sub-optimal, because «although there will be occasions on which the rule-indicated result will be inferior to the justification-indicated result, there will be no occasions on which the rule-indicated result will be superior to the justification-indicated result»[91]. But this does not prove that the adoption of rule-based decision-making is irrational. We must distinguish between the justification or purpose behind the rule, i.e., the "substantive justification", and the justification for *having a rule* as a strategy to serve the substantive justification – the "rule-generating justification", the justification for specifying the (substantive) justification in the form of a rule[92].

[84] RAZ 1999 [1975, 1990], 49.

[85] SCHAUER 1991, 10, 168 ff. Schauer contrasts rules imposing substantive constraints with *naked jurisdictional rules*, and argues that while a legal system can dispense with the former—as exemplified by the Weberian ideal of *qadi* justice—it cannot function without the latter.

[86] WALDRON 2016, parag.4

[87] SCHAUER 1991, 129.

[88] SCHAUER 1991, 26

[89] SCHAUER 1991, 29

[90] SCHAUER 1991, 31-37.

[91] SCHAUER 1991, 100.

[92] SCHAUER 1991, 94.

As we said, the use of rules to guide people's conduct is only one among possible strategies. Alternatively, the agent may be left – in a particularistic fashion – to contemplate the balance of background reasons which would do all the normative work without the interposition of the rule[93]. Although intuitively superior, a decision-making procedure that aims to produce the best outcome for each case may be self-defeating, for instance because it wastes decision-making resources or is prone to error. The need to spare decisional resources and the distrust of a certain category of decision-makers – as well as other similar considerations – are the justification for the adoption of rules, which, in general, are tools for the allocation of decisional power[94]. This is the justification for having rules and this is why, rule-based decision-making may overall be the best (the most rational) decision-making procedure available[95].

That said, since rules are crucial in our societies, it is appropriate to emphasize the role of rule-based decision-making within the Service Conception of Authority, talking of the Service Conception *of Rules*, even though the considerations that will follow are relevant for obedience to authoritative directives in general. Accordingly, the exclusionary reasons under analysis will be authoritative rules, which – following the Razian framework – are to be understood as second-order reasons for not being motivated by those first-order reasons that fall within the scope of the rule[96].

The distinction between, on the one hand, the strategies that agents should follow in their deliberation – the second, derived, level – and, on the other, the principles or theories that provide the ultimate criteria of correctness, and yet instruct people not to be motivated by them directly – the first level –, can be used to compare indirect consequentialism and the Service Conception of Rules. We arrive at the following scheme:

> First level: Consequentialism / the substantive justification of the rules (in general, the theory about background reasons)
>
> Second level: Improved version of common-sense morality / rule-based decision-making (in general, the Exclusionary Model of decision-making)

Indirect consequentialism is the conjunction of consequentialism as the ultimate criterion of correctness and, roughly, an improved version of common-sense morality as the model of decision-making. While the Service Conception of Rules is an indirect theory of decision-making which prescribes adopting a rule-based model of decision-making to serve better the balance of background reasons.

While both are indirect theories of decision-making, the Service Conception of Rules and indirect consequentialism differ in two important and intertwined aspects.

The first difference regards the relation with our intuitions. In the indirect consequentialist model, consequentialism operates at the first level – being the ultimate theory about the criteria of correctness –, while common-sense morality is adopted as the model of decision-making. Consequentialism is not an intuitive method of decision-making, while *common-sense* morality by definition respects people's pre-theoretical moral intuitions; therefore, the practical result of indirect consequentialism is, in many cases, the legitimation of people's common intuitions. Indeed, although Parfit's analysis reveals that moral truths are complex and counterintuitive (if

[93] SCHAUER 1991, 94.

[94] SCHAUER 1991, ch. 7, see 158 ff. in particular.

[95] SCHAUER 1991, 102. I am skipping here Schauer's important discussion about "rule-sensitive particularism" (SCHAUER 1991, 97).

[96] Schauer does not consider rules as exclusionary reasons in the same way as Raz does, but we can neglect this difference (see SCHAUER 1991, 88 ff.).

consequentialism is true, not only we have counterintuitive duties[97], but we also must accept that moral truths may be self-effacing and esoteric), it nevertheless reaffirms the value of common-sense intuitions as a potentially[98] legitimate method of decision-making. In a nutshell: a complex, disconcerting, counter-intuitive theory of practical reasons legitimizes a simpler and more intuitive decision-making model. This pattern is reversed in the case of the Service Conception of Rules (and generally in the case Service Conception of Authority which is applied to the broader category of authoritative directives). In a sense weighing all the practical reasons is more complex than weighing just the unexcluded reasons. However, it is also true that from the agent's perspective the substitution is not between balancing reasons and balancing non-excluded reasons, but between one's own sense of justice and obedience to the rule. Although from the normative standpoint the shift from the first level to the second is a simplification of the task, from the phenomenological and psychological standpoint the shift is rather a complication, at least until we rule out factors such as the interiorization of rules. In other words, while in the scheme of indirect consequentialism the complex is substituted with the simple, in the case of the Service Conception of Rules the simple – illusorily simple, but still psychologically simple – is substituted with the complex. Rule-based decision-making and in general self-conscious obedience to authoritative directives is rational because it is a safer procedure of decision-making, but is cognitively laborious, because the agent is required to neutralize her own moral intuitions, and to accept counter-intuitive conclusions.

Speaking about rules in particular (but the point can be applied to particular authoritative directives as well) Schauer says:

> «the authority makes in advance the best all things considered decision (...) by creating a procedure in which the subject will be dissuaded by exercising that subject's *best* (but expected to be mistaken) judgment (...) then the rule-maker's task is often one of inducing a rule-applier or rule-subject to relinquish her best judgment»[99] (italics added).

As anticipated, there is also another important difference between indirect consequentialism and the Service Conception of Rules, which reinforces the first one. As we have seen, Parfit leaves open the possibility that consequentialism may instruct many people to believe in another moral theory: consequentialism may be partially self-effacing and partially esoteric. This solution can hardly be used in the context of legal rules and authoritative decisions, as far as liberal democracies are concerned.

First of all, it is significant that Parfit, in analysing how self-interest theory and consequentialism may be self-effacing and esoteric, resorts to thought experiments involving drug use or phenomena such as hypnosis[100]. These solutions are clearly unavailable in the context of the Rule of Law. We may add that in order to conceal consequentialism there is no need, in the actual world, to resort to drugs and hypnosis, because the consequentialist way of reasoning is not common among people. Instead, reference to the substantive justification of authoritative rules is difficult to erase from people's minds, since it is grounded on their sense of justice. What is often neglected by people is not the substantive justification of the rule, but rather the justification for

[97] The same does not apply to self-interest theory, which may be regarded as an intuitive theory of rationality by many. This is why I am confining the analysis to consequentialism.

[98] The adoption of a revised version of common-sense morality would be justified if there were a risk that too many people might adopt consequentialism. A scenario far removed from present reality, according to Parfit, as we noted above.

[99] SCHAUER 1991, 133 f.

[100] PARFIT 1984, 12 f., 41.

having a rule. While consequentialism spontaneously tends to be self-effacing and esoteric, the untutored, naïve, sense of justice of people is by definition a common, public, mindset.

Therefore, unless *ad hoc* social techniques aimed at changing people's sense of justice intervene, the second level will not cognitively replace the first, but rather will be added to it. Not only must the agent change her intuitive method of decision-making, but she also *remembers* that the counter-intuitive method she must follow is a sub-optimal one. Authoritative rules may not only demand actions that are really wrong, but also do so in full view of the agent's moral *awareness*. The Service Conception of Rules has, in this sense, a "genealogical" nature. This intensifies the cognitive load on the obedient agent, whose sense of justice is in direct conflict with the content of the authoritative directive and cannot simply be set aside. The agent experiences that the truth about practical are "depressingly convoluted"[101].

The Service Conception of Rules is an indirect theory which both (i) legitimizes a counter-intuitive method of decision-making and (ii) tends to be genealogical (rather than self-effacing and esoteric).

7. *Conclusions*

In the indirect theories of decision-making reasoning must worsen in order to improve. Indirect consequentialism and the Service Conception are both indirect theories of decision-making. Rule-based decision-making – which is at the core of the Rule of Law – is one of the most important decision-making procedures which are justified by the Service Conception. The comparison between indirect consequentialism and the Service Conception shows a deep structural difference. On the one hand, there are indirect theories that legitimize intuitive decision-making procedures and are self-effacing; on the other hand, there are indirect theories that legitimize counter-intuitive decision-making procedures and have a genealogical character. When indirect consequentialism tells people not to be motivated by consequentialism but by an improved version of common-sense morality the "worsening" of reasoning is not felt by the agent, whereas the worsening is felt when the Service Conception of Authority tells people not to be motivated by their assessment of the background reasons but by authoritative directives. The agent subjected to authoritative directives will find right things that are contrary to the directives and wrong things that are commanded by the directives; moreover, she will know that the criteria she is using are suboptimal, and then that some actions contrary to the directive that seem right are really right, and conversely that some actions prescribed by the directive that seem wrong are really wrong. This may occur either because the legitimate authority has issued a wrong directive, or because the directive is a rule which, although abstractly correct, commands a wrong action in the specific case. Acknowledging this requires control, control is effortful, and effortful things may look wrong. As Dennett noted when talking about the idea of competence without comprehension: «Our skepticism about competence without comprehension has causes, not reasons. It doesn't "stand to reason" that there cannot be competence without comprehension; it just feels right, and it feels right *because* our minds have been shaped to think that way»[102]. Rule-based decision-making is part of an unfortunate kind of indirect theories of reasoning.

[101] Parfit 1984, 23.
[102] Dennett 2017, 58.

Bibliographical references

ADAMS R.M. 1976. *Motive Utilitarianism,* in «The Journal of Philosophy», 73, 14, 1977, 467 ff.

ADAMS N.P. 2021. *In Defense of Exclusionary Reasons*, in «Philosophical Studies», 178, 1, 2021, 235 ff.

RUTH C. 2016. *Comparativism: The Grounds of Rational Choice*, in LORD E., MAGUIRE B. (eds), *Weighing Reasons*, Oxford University Press, 213 ff.

COCKING D., OAKLEY J. 1995. *Indirect Consequentialism, Friendship, and the Problem of Alienation*, in «Ethics», 106, No. 1, 1995, 86 ff.

DENNETT D. C. 2017. *From Bacteria to Bach and Back. The Evolution of Minds*, Norton & Company.

GRIFFIN J. 1986. *Well-being. Its Meaning, Measurement, and Moral Importance*, Oxford University Press.

GUR N. 2018. *Legal Directives and Practical Reasons*, Oxford University Press.

JORDAN A. 2018. *Exclusionary Reasons, Virtuous Motivation, and Legal Authority*, in «Canadian Journal of Law & Jurisprudence», XXXI, No. 2, 2018, 347 ff.

MACKIE J. L. 1977 *Ethics: Inventing Right and Wrong,* Penguin books.

MCNAUGHTON D., RAWLING P. 2025. *Consequentialism*, available at https://www.rep.routledge.com/articles/thematic/consequentialism/v-1.

MOORE, M. S. (1989). *Authority, law, and razian reasons*, in «Southern California Law Review», 62, 4, 1989, 827 ff.

MUFFATO N. 2022. *Ragioni esclusive, concezione gerarchica della razionalità e precetti autoritativi*, in «Ragion Pratica», 59, 2, 2022, 503 ff.

NOZICK R. 1975. *Anarchy, State, and Utopia*, Blackwell Publishers.

PORTMORE D.W. 2001. *Can an Act-Consequentialist Theory Be Agent Relative?*, in «American Philosophical Quarterly», 38, 4, 2001, 363 ff.

PARFIT D. 1984. *Reasons and Persons*, Oxford University Press.

RAILTON P. 1988. *Alienation, Consequentialism, and the Demands of Morality*, in SCHEFFLER S. (ed.), *Consequentialism and Its Critics*, Oxford University Press, 93 ff.

RAZ J. 1986. *The Morality of Freedom*, Oxford University Press.

RAZ J. 1989. *Facing Up: A Reply*, in «Southern California Law Review», 62, 1989, 1153 ff.

RAZ J. 1999. *Practical Reasons and Norms*, Oxford University Press (ed. or. 1975; revised edition with postscript 1990).

RAZ J. 2006. *The Problem of Authority: Revisiting the Service Conception*, in «Minnesota Law Review», 90, 2006, 1003 ff.

RAZ J. 2009. *The Claims of Law*, in ID., *The Authority of Law* (second edition), Oxford University Press, 28 ff.

SCHAUER F. 1991. *Playing by the Rules. A Philosophical Examination of Rule-based Decision-Making in Law and in Life*, Oxford University Press.

SEN A.K., WILLIAMS B.A.O. 1982, *Utilitarianism and Beyond*, Cambridge University Press.

SMEULERS A. 2019. *Why Serious International Crimes Might Not Seem 'Manifestly Unlawful' to Low-level Perpetrators: A Social–Psychological Approach to Superior Orders*, in «Journal of International Criminal Justice», 17, 1, 2019, 105 ff.

SMITH M. 1994. *The Moral Problem*, Blackwell Publishing.

VASSILIOU A. 2022. *The Normativity of Law: Has the Dispositional Model Solved our Problem?* in «Oxford Journal of Legal Studies», 42, 3, 2022, 943 ff.

WALDRON J. 2016. The Rule of Law, in «Stanford Encyclopedia of Philosophy», available at https://plato.stanford.edu/entries/rule-of-law/#:~:text=The%20Rule%20of%20Law%20comprises,norms%20that%20govern%20a%20society.

WHITING D. 2017. *Against Second-Order Reasons*, in «Noûs», 51, 2, 2017, 398 ff.

WILAND E. 2007. How Indirect Can Indirect Consequentialism Be?, in «Philosophy and Phenomenological Research», 74, 2, 2007, 275 ff.

WILLIAMS B.A.O. 1973. *A Critique of Utilitarianism*, in SMART J.J.C., WILLIAMS B.A.O. (eds.), *Utilitarianism For & Against*, Cambridge University Press, 77 ff.

WILLIAMS B.A.O. 1981. *Persons, Character and Morality*, in Id., *Moral Luck. Philosophical Papers 1973-1980*, Cambridge University Press, 1 ff. (ed. or. 1975).

WILLIAMS B.A.O. 2006. *Ethics and the Limits of Philosophy*, Routledge (ed. or. 1985).