Political Theory
Christian List and Laura Valentini[*]

Political theory, sometimes also called "normative political theory", is a subfield of the disciplines of philosophy and political science that addresses conceptual, normative, and evaluative questions concerning politics and society, broadly construed. Examples are: When is a society just? What does it mean for its members to be free? When is one distribution of goods socially preferable to another? What makes a political authority legitimate? How should we make collective decisions? What goals should our policies promote? How should we trade off different values, such as liberty, prosperity, and security, against one another? What do we owe, not just to our fellow citizens, but to people in the world at large? Is it permissible to buy natural resources from authoritarian governments? Can war ever be just?

Political theory is a long-established field. Its questions have animated thinkers since Classical Antiquity. Consequently, the methods of theorizing and substantive conclusions are diverse. In this article, we review the methodology of a core branch of contemporary political theory: the one commonly described as "analytic" political theory.

Given space constraints, we are not able to cover the history of political thought, the study of ideologies, the comparative study of political thought across cultures, and "continental" political theory, including "hermeneutic", "post-structuralist", and "post-modernist" approaches (for a more comprehensive discussion, see Leopold and Stears 2008). Nonetheless, the label "analytic" should not be interpreted narrowly. It is meant to refer to an argument-based and issue-oriented, rather than thinker-based and exegetical, approach that emphasizes logical rigour, terminological precision, and clear exposition. The term "analytic" is not intended to refer only to the logical and linguistic traditions of philosophy associated with the Vienna Circle and philosophers in Oxford and Cambridge in the first half of the 20th century. Contemporary analytic political theory goes well beyond conceptual analysis. At least since the publication of John Rawls's *Theory of Justice* (1971/1999), the development of normative *theories* has been one of the field's central concerns.

This review is structured as follows. In Section 1, we briefly demarcate the scope of political theory. In Section 2, we comment on the analysis of political concepts. In Section 3, we introduce the notions of principles and theories, as distinct from concepts. In Section 4, we discuss the methods of assessing such principles and theories, for the purpose of justifying or criticizing them. In Section 5, we review a recent debate on how abstract and idealized political theory should be. In Section 6, finally, we discuss the significance of disagreement in political theory.

One clarification about the nature of this review is needed. Although we cover established ground, we do so from an angle that will be somewhat unfamiliar to at least some political theorists – namely an angle inspired by the philosophy of science. We have chosen this angle with a view to systematizing the activity of analytic political theorizing so as to make its connections with other fields of philosophy and

positive science more transparent. This seems appropriate in the context of a handbook on philosophical methodology.

1. The scope of political theory

To demarcate the scope of political theory, it is helpful to distinguish it from its most closely related neighbouring fields: political science, moral philosophy, legal theory, normative economics, and social ontology. We also offer some comments on the use of the label "political theory" as opposed to "political philosophy".

1.1 Political theory and political science

Political theory can easily be distinguished from (positive) political science. Political science addresses empirical and positive questions concerning politics and society (for an overview, see Goodin 2009). It seeks to describe and explain actual political phenomena, such as elections and electoral systems, voter behaviour, political-opinion formation, legislative and governmental behaviour, the interaction between the legislative, executive, and judicial branches of the state, and the stability or instability of different forms of government. Political theory, by contrast, addresses conceptual, normative, and evaluative questions, such as what a democracy is, how we ought to organize our political systems, and how to evaluate the desirability of policies.

Political theory and political science can complement each other. Normative recommendations and evaluations of policies or institutional arrangements often rest on empirical premises. It is hard to arrive at a blueprint for a just society, for example, without understanding how society actually works, since normative recommendations may have to respect feasibility constraints (e.g., Gilabert and Lawford-Smith 2012). Thus political theory requires political science and the social sciences more generally. Similarly, when political scientists investigate, for instance, whether democracy promotes economic development or whether free societies are more politically stable and less corrupt than unfree ones, they need to know what counts as a democracy or how to define freedom. These questions require the conceptual input of political theorists. Finally, among the large number of empirical questions that political scientists could investigate, some are undoubtedly more interesting, relevant, and pressing than others, and political theory can help shape the research agenda.

1.2 Political theory and moral philosophy

While there is a natural division of labour between political theory and political science, the distinction between political theory and moral philosophy is subtler. Some scholars view political theory as a subfield of moral philosophy, in which the concepts and principles from moral philosophy are applied to political – and, more broadly, social – problems. If one views political theory in this way but also does not want it to collapse completely into moral philosophy, one must give a clear criterion of when a problem counts as "political".

Unfortunately, this is not straightforward. For instance, saying that a problem is political if it involves multiple people or their living together may seem plausible, but is too inclusive. Many problems in personal ethics, such as how to treat one's

relatives or friends, would count as political on this criterion, even though they are not generally classified in this way. On the other hand, saying that a problem is political if it involves the state or government is too restrictive, because a number of problems outside the sphere of the state or government narrowly construed might still be seen as political.

Consider, for instance, gender relations in civic life or in the workplace, which many people regard as a political issue. The question of whether, and how, a distinction between the private or personal domain and the political or social one can be meaningfully drawn is controversial. Some feminist thinkers have famously challenged the possibility of drawing any such distinction and have endorsed the slogan "the personal is political" (for an overview, see Baehr 2013). In sum, if one wanted to distinguish political theory from moral philosophy by referring solely to the *substantive domain* of problems addressed, one could at best use some heuristic criterion to capture a conventional distinction, but this would yield no principled line.

Another way to distinguish political theory from moral philosophy is to invoke the *conditions of theorizing* in each field. The aim of moral theorizing, one might say, is to come up with the correct solution to any given moral problem *simpliciter* – the solution that, based on the theorist's comprehensive moral view, turns out to be right. The aim of political theory, one might argue, is different. The political theorist, at least under modern conditions, is engaged in problem solving under a particular constraint: the presence of pluralism and disagreement about how to solve the problem at hand (see Rawls 1996; Waldron 1999). Thus any compelling solution to the problems of political theory, such as how to define justice or how to design a legitimate procedure for collective decision-making, must appeal to people with a variety of (reasonable) viewpoints, precisely because those solutions are meant to apply to, and be acceptable in, pluralistic societies.

If we use such a methodological, rather than substantive, criterion for distinguishing political theory from moral philosophy, we need not worry about identifying a particular domain of problems that count as political. Rather, we can say that the hallmark of political theory is its mode of theorizing, against the background of (reasonable) pluralism. We return to this idea in Section 6, where we discuss the significance of disagreement in political theory. Of course, substantive and methodological criteria for distinguishing political theory from moral philosophy can be combined.

1.3 Political theory and legal theory

It is useful to comment briefly on the relationship between political theory and legal theory. The two overlap, and it is difficult to draw a sharp distinction between them. We might again arrive at a rough distinction by using some heuristic criterion to identify what counts as "legal" or "related to the law". As with the attempt to distinguish the "political" from the "private" or "personal", however, we cannot expect any such criterion to yield a definitive distinction.

Alternatively, we might try to distinguish political theory from legal theory by identifying different modes of theorizing associated with each field. For instance, an argument to the effect that justice *simpliciter* requires respect for human rights and

certain universal welfare protections is distinct from an argument to the effect that a particular constitution or kind of legal system, properly interpreted, requires them. One might say that the former argument belongs to political theory, the latter to legal theory (see the discussion of the nature of legal interpretation in Dworkin 1986).

More generally, one might say that the deontic concepts used in legal theory, such as *legal* permissibility, *legal* rights, and *legal* obligations, are different from their counterparts in political or moral theory and therefore require a different analysis.[1] Still, political theory and legal theory are best seen as overlapping fields of enquiry.

1.4 Political theory and normative economics

Political theory also overlaps with normative economics, especially with social choice and welfare theory. Social choice and welfare theory is the formal, but also normative and evaluative, study of (i) collective decision-making procedures, (ii) mechanisms for allocating benefits and burdens in society, and (iii) methods by which a social planner, policy maker, or institutional designer can assess the goodness or desirability of different social states, policies, or institutions. Normative economists investigate these – (i), (ii), and (iii) – by introducing desiderata that any reasonable procedure, mechanism, or method is required to satisfy and then asking which procedures, mechanisms, or methods, if any, satisfy the given desiderata. (For a survey, see List 2013.)

The substantive questions addressed in social choice and welfare theory are similar to some of those addressed in political theory. For example, questions such as how to distribute benefits and burdens are addressed by political theorists and normative economists alike. Indeed, Rawls's *Theory of Justice* was, in part, influenced by the normative works of economists such as Kenneth Arrow (1951), John Harsanyi (1955), and Amartya Sen (1970). Similarly, normative economists frequently draw on moral and political theory. For example, John Roemer's (1998) formal work on distributive justice is influenced by G. A. Cohen's (1995) work in political theory; and the work on variable-population social choice by Charles Blackorby, Walter Bossert, and David Donaldson (2005) is influenced by Derek Parfit's (1984) work on population ethics.

Arguably, the main difference between political theory and social choice and welfare theory is not a substantive one (although the former is broader than the latter), but a methodological one. Mainstream political theory is a non-formal discipline, making at most limited use of formal methods from mathematics, logic, and economics, while social choice and welfare theory is predominantly formal.

---

[1] Of course, even within legal theory, one can distinguish between what is legally permissible according to a specific set of laws (e.g., those of a particular country) and what the standard of legal permissibility should be.

1.5 Political theory and social ontology

A less well-known but growing field in the neighbourhood of political theory is social ontology. Social ontology investigates the nature of phenomena such as joint intentions, collective actions, social norms and conventions, group agency, and institutions (e.g., Searle 1995; Bratman 1999; Gilbert 2006; Tuomela 2007; Tollefsen 2002; List and Pettit 2011). Its central question, roughly speaking, is: What are the building blocks of the social world, and how are they related to one another, to the individuals involved, and to the physical world?

The substantive questions in social ontology are distinct from those in political theory and in some respects prior to them. Social ontology is primarily a positive and explanatory field rather than a normative or evaluative one. Its relevance to political theory lies in the fact that political theory cannot get off the ground unless we are clear about what entities and properties exist in the social world. For example, before we can answer questions about rights, obligations, and responsibilities, we need to know what entities can be the bearers of rights, obligations, and responsibilities. Are these restricted to individuals, or do they also include certain collectives? Should we regard the state as an agent, as a mere collection of individuals, or as some kind of special fiction? Similarly, before we can answer the question of what is or is not socially desirable, we need to know what the possible objects of value might be. It is difficult to determine, for instance, whether there should be any cultural rights or state subsidies for cultural activities unless we give can give at least a partial answer to the question of what we mean by "culture".

Since different social-ontological theories give us different accounts of what entities and properties there are in the social world, they can, in turn, impose constraints on what the possible objects of analysis in political theory might be.

1.6 Political theory and political philosophy

Although some scholars distinguish between political *theory* and political *philosophy*, this is mainly a conventional distinction. It refers, roughly, to the different styles of political theory conducted in political science and philosophy departments, respectively, especially in North America. (In the UK, much of what is conventionally called "political philosophy" is traditionally conducted in political science departments.) Arguably, "political theory" is the slightly more inclusive and interdisciplinary label, referring not only to philosophical work but also to a variety of other approaches. As mentioned, here we focus specifically on the analytic branch of political theory.

2. Concepts in political theory

A long-standing concern in political theory is the analysis of political concepts: *freedom*, *equality*, *justice*, *authority*, *legitimacy*, *democracy*, *welfare*, and so on. Each of these has been interpreted and defined in numerous ways, and political theory can help us clarify the advantages and disadvantages of different interpretations and

definitions. The bulk of political theory in the decades prior to the publication of Rawls's *A Theory of Justice* was conceptual analysis.[2]

The analysis of political concepts is relevant not only to normative theory building (e.g., any *theory* of liberalism must employ some *concept* of liberty) but also, as already noted, to positive work in political science. Thus conceptual analysis is, in some sense, the least normative or evaluative part of political theory.

Of course, when we analyze concepts such as *freedom* and *democracy*, we are usually interested in the kinds of freedom and democracy that we find valuable or normatively required. Hence the ultimate motivation for our analysis may well be a normative or evaluative one: we may wish to clarify these concepts in order to express normative or evaluative principles in terms of them. Logically, however, the question of how to understand *freedom* and *democracy* is distinct from the question of whether freedom and democracy so understood are valuable. Indeed, political scientists may sometimes be interested primarily in whether *freedom* or *democracy* can serve as independent variables in explanations of political phenomena (e.g., when they investigate whether freedom promotes prosperity or whether democracies are less prone to fighting wars against one another), irrespective of any considerations of value.[3] In this section, we introduce some key ideas relevant to the analysis of political concepts.

2.1 What is a concept?

We use concepts to categorize or classify objects.[4] The concept *democracy*, for example, may help us distinguish between those forms of government that are democratic and those that are not. The concept *legitimacy*, in relation to acts of state coercion, may help us distinguish between those acts of state coercion that are legitimate and those that are not.

For the purposes of this article, we assume that any concept has a *domain of application*. This is the set of objects of which it is meaningful to ask whether they fall under the given concept or not. We might say, for example, that the domain of the concept *democracy* is the set of all systems of government or decision-making. For any object in that set – i.e., any system of government or decision-making – we can meaningfully ask whether it is democratic or not. By contrast, for objects outside that domain, it is not meaningful to ask whether they are democratic or not. It makes no sense to ask, for instance, whether an equilateral triangle or a mountain is democratic or not (cf. Dworkin 1986, 75). Note that the domain of the concept *democracy* – or

---

[2] For a classic work on political concepts, see Oppenheim (1981).

[3] Of course, political scientists may also consider *freedom* or *democracy* as dependent variables, such as when they ask which social and economic conditions promote each. Think, for instance, of the literature on the transition to, and consolidation of, democracy (e.g., Linz and Stepan 1996). In such studies, considerations of value may plausibly affect our choice of interpretation of the relevant concepts.

[4] We here cannot discuss the ontological status of concepts, on which there are several rival views in philosophy. Instead, we rely on a relatively simple characterization of concepts, emphasizing the fact that we use concepts to categorize or classify objects and that they serve as ingredients in the activity of political theorizing. For a broader discussion of concepts as the constituents of thoughts, see Margolis and Laurence (2012). For a cognitive-science treatment of concepts as locations or regions within "conceptual spaces", see Gärdenfors (2000).

justice, or *freedom*, and so on – can be variously specified; we return to that point below.

Second, any "classically well-behaved" concept has *defining conditions*. These determine, for any object in the concept's domain, whether that object falls under the concept ("satisfies it") or not. In our example, the question of whether a particular system of government or decision-making, say the political system of Iceland, is democratic or not depends on a variety of features of that system: how decisions are made, who participates in those decisions, how the participants provide their input, how the decisions are implemented and by whom, and so on.

Philosophers are divided over the extent to which concepts in general have defining conditions (for a review, see, e.g., Margolis and Laurence 2012). Many of our common-sense concepts arguably lack such conditions. We may be able to pick out some paradigm instances (or "prototypes") of *redness* or *beauty*, but may be unable to arrive at clear-cut necessary and sufficient conditions that an object must satisfy to be red or beautiful (for a related discussion, see Dworkin 1986, chap. 2). By contrast, in theoretical work, it is usually desirable to look for concepts with defining conditions. Occasionally, however, some theoretical concepts may be regarded as undefined "primitives" or as characterizable only through "prototypes".

Finally, a concept's *extension* is the subset of the domain consisting of precisely those objects that fall under the concept ("that satisfy it"). If the concept has defining conditions, these determine the extension. The extension of the concept *democracy* is the set of all those systems of government or decision-making that, according to the concept, count as democratic. More generally, there can be "non-binary" concepts, which do not subdivide objects into those falling under the concept and those not falling under it, but which instead classify objects on one or several dimensions that may each admit of degrees. For example, *equality* and *welfare* are non-binary concepts. The level of equality or inequality in a particular distribution of goods is a matter of degree, and different interpretations of *equality* give us different accounts of when one distribution counts as more equal than another (see, e.g., Sen 1980). Similarly, a person's welfare is a matter of degree and, on some accounts, even given by a vector of multiple attributes, representing different dimensions or aspects of welfare. Sen (1987) has argued for the "constitutive plurality" of the concept *standard of living*; there are multiple dimensions on which a person's standard of living can be categorized.

For practically any salient concept in political theory, there are debates about what the domain of application is, what the defining conditions, if any, are, and which objects belong to the concept's extension and which do not. Just think of the many different ways in which the concept *democracy* may be understood. The domain of application may be specified in a variety of ways: for example, as a set of voting procedures, as a set of decisions, or as a set of entire political systems. Similarly, we may be divided over the defining conditions that determine whether something is democratic or not: for example, do only formal, constitutional features of the political system matter, or are features of actual political practices relevant as well, and if so, how? And is democracy precisely definable at all? In consequence, we may end up with different extensions of the concept *democracy* (see, e.g., Christiano 2008; List 2011). Similar considerations also hold for other key political concepts, such as *justice*, *freedom*,

equality, and *legitimacy*. Indeed, many political concepts are what political theorists call "essentially contested" (Gallie 1955).

2.2 Concepts and conceptions

Political theorists, following Rawls, who in turn follows H.L.A. Hart, sometimes distinguish between *concepts* and *conceptions* (Rawls 1971/1999, 5). Concepts, in that terminology, are less fully specified than conceptions. For example, we may have a broadly outlined *concept* of freedom as the absence of constraints on agents' actions, which still leaves open what kinds of constraints, agents, and actions matter (MacCallum 1967). A full specification of those constraints, agents, and actions yields a precise *conception* of freedom. Different conceptions can thus be compatible with the same broad concept.

We can translate the distinction between concepts and conceptions into our earlier terminology by defining a *conception* exactly as we defined a concept in the last section, where the domain, defining conditions (if any), and extension are fully specified, and redefining a *concept* as a broader family of such conceptions, with some aspects of the domain, defining conditions, or extension left open.

2.3 Desiderata on concepts

The following is a list of desiderata that systematize requirements often implicitly employed by political theorists engaging in conceptual analysis:

Respecting our intuitions: We may want to interpret a concept, such as *freedom* or *democracy*, in a way that is broadly in line with our intuitions, especially if this concept has a common-sense interpretation. If we arrived at an interpretation of *freedom* that classifies some intuitively clear cases of unfreedom as instances of freedom, or vice versa, this would be suspect. For example, G.A. Cohen has famously criticized Robert Nozick's "moralized" conception of freedom (according to which freedom is, roughly, the absence of rights-violating interference) on the grounds that it delivers counterintuitive judgments, such as that a *justly* imprisoned criminal is not unfree (G. A. Cohen 1988). Similarly, we would be disturbed if our interpretation of *democracy* classified some clear-cut cases of tyranny as democratic. In such cases, we would either have to give strong reasons for overruling our intuitive judgments or search for a better interpretation of the concept. Later, we discuss Rawls's method of *reflective equilibrium*, which may help us adjudicate cases in which our theoretical conclusions conflict with our intuitive judgments (Rawls 1971/1999; Daniels 2013).

Playing the right normative, evaluative, or descriptive role: We may be interested in a concept because we wish to use it in some normative or evaluative principle or in some explanation in political science. A good interpretation of the concept is one that successfully plays the intended normative, evaluative, or descriptive role. Different roles might require different interpretations of the concept. For example, if we want the concept of justice to offer a comprehensive picture of how society should be organized, we are likely to develop a thicker account of it than if we understand justice as one value among the many that should guide institutional design (Rawls 1971/1999; G. A. Cohen 2008, 271–272). However, to avoid a proliferation of rival

interpretations of the same concept, we might also be looking for a single interpretation that can successfully play multiple roles.

Standing in the right relationship to other concepts: Since we typically employ concepts not in isolation but in connection with other concepts, we may require these concepts to be related to each other in the right way. If we take rights to entail obligations, for example, our interpretation of *rights* may constrain the way in which we can consistently interpret *obligations*. Sometimes we may wish some concepts to be directly inter-definable. It is often held, for instance, that the concept *permissibility* must be definable in terms of the corresponding concept *obligatoriness* and vice-versa: it is permissible that *p* if and only if it is not obligatory that *not p*. Consequently, our joint analysis of *permissibility* and *obligatoriness* must respect this constraint. Finally, we may wish some given concepts to be sufficiently "differentiated" from one another, in order to avoid redundancies or confusions. (For a related discussion of concept formation in the social sciences, see Gerring 1999.)

Having defining conditions that are neither too "thick" nor too "thin": Even when we have settled the domain of a given concept (e.g., we focus on a concept of freedom whose domain is a set of acts, as opposed to a set of agents or institutional arrangements), we might still be divided over the concept's defining conditions. In the case of *freedom*, a huge variety of different defining conditions have been proposed (for overviews, see Carter 2012; Lovett 2013). To narrow down the range of possibilities, we might require that the defining conditions be neither too "thick", nor too "thin": they should not refer to any "irrelevant" facts about the objects to be categorized, but refer to all "relevant" facts. There can then be debates about which facts are or are not relevant and what counts as too "thick" or too "thin". For example, a concept is *moralized* if its defining conditions refer to some normative or evaluative facts. A concept is *non-moralized* otherwise.[5] Nozick's above-mentioned concept of freedom as the absence of rights-violating interference is moralized in this sense. Political theorists are often divided over which concepts in political theory should be moralized. Similarly, a concept is *modally demanding* if its defining conditions refer not only to facts about the actual world but also to facts about other possible worlds – i.e., facts about what would be or might be, not merely about what is. A concept is *modally undemanding* otherwise. Pettit, for instance, argues that *freedom* is modally demanding: a slave with a benevolent, non-interfering master still counts as unfree, because there is a nearby possible world in which the master interferes (Pettit 1997). Similarly, concepts such as *security* and *peace* are arguably modally demanding, referring not only to the absence of relevant harmful actions or military conflicts in the actual world but also their continued absence in a range of nearby possible worlds. Political theorists are divided over which, if any, concepts should be modally demanding (see, e.g., Pettit 2011; Southwood 2013).

Having defining conditions that are epistemically accessible: Depending on the intended use of a concept, we may require its defining conditions to be such that it is possible, at least in principle, for us to *know* whether an object meets them. For example, a concept of welfare whose defining conditions refer to certain kinds of mental states that are inaccessible to any observer would be of little practical use.

---

[5] The definition of moralized and non-moralized concepts requires suitable adjustments if there are no evaluative or normative facts.

Similarly, the defining conditions of *justice* under Robert Nozick's account, which refer to the entire history of transactions leading to the current distribution of entitlements, fail to meet this desideratum and thus render Nozick's account somewhat practically inert (Nozick 1974). Of course, the context and intended use may determine what counts as epistemically accessible.

2.4 What concepts are not

Concepts should not be confused with principles or theories – the topic of Sections 3 and 4. In particular, principles and theories have propositional content and may be true or false, while concepts, by themselves, cannot be true or false. They only categorize objects and, in doing so, can be more or less useful, more or less plausible, and more or less in line with established use or with our considered judgments.[6] To give a simple illustration from outside political theory, the concepts *red*, *green*, or *triangular* are neither true nor false. Only statements in which they occur can have truth-values, such as "tomatoes are red" (true), "snow is green" (false), and "rectangles are triangular" (false).

Even if we have a full account of when a political system is democratic or what it means for someone to be free, this still leaves open the question of whether democracy or freedom are desirable and whether we ought to promote each of them. We need principles or theories – making statements such as "we ought to respect freedom" or "we ought to make decisions democratically" – to address the latter questions. Such principles or theories are then capable of being true or false. Simply put, concepts can serve as building blocks of principles, which can serve as building blocks of theories.

Still, people sometimes say things such as "freedom as non-interference is the true conception of freedom". Assertions of this kind are best interpreted as abbreviations for claims such as "freedom as non-interference is the conception of freedom that is, in some relevant sense, most appropriate", which, in turn, could mean that it best captures our established use of *freedom* or alternatively our considered judgments about what counts as free. Literally, however, the claim that one concept of freedom is the true one is not meaningful, since only things with propositional content can have truth-values. Concepts do not have propositional content: the extension of a concept is not a set of possible worlds (which is what the extension of a proposition is), but a set of objects.[7]

Even on the above-mentioned charitable interpretation in terms of appropriateness, the assertion "freedom as non-interference is the true conception of freedom" is ambiguous. As noted, it would have to be understood as implying that, among the many different interpretations of *freedom*, one stands out as "most appropriate". But, unless we specify a criterion of appropriateness, there is no unique such interpretation. For example, one interpretation may be most in line with our common-sense use of the word "freedom", another most in line with our considered judgments about what counts as free, a third most suitable for playing a particular role in a theory

---

[6] Recall that the defining conditions of any concept simply specify when an object falls under the concept, not whether the concept is true or false.

[7] Note that the extension of a statement or proposition is the set of those possible worlds in which the statement or proposition is true.

of justice. Different criteria of appropriateness may diverge, and there is no application-independent criterion. Thus expressions such as "the true conception of freedom" are, at best, shorthand for more elaborate expressions involving criteria of appropriateness. To avoid ambiguity, it is best to spell those out explicitly.

3. Principles and theories in political theory

While analytic political theory until the early 1970s was primarily concerned with the analysis of concepts, John Rawls's *Theory of Justice* (1971/1999) invigorated the quest for theories and the principles underlying them. Rawls formulated some principles of justice, which are the basis of his theory of how we should organize the "basic structure of society", namely its main political, legal, and economic institutions.

In this section, we discuss the notions of principles and theories. Although these are widely used in political theory, they are seldom carefully defined. We hope, therefore, that our discussion will be clarifying.

3.1 What is a principle?

A *principle* is a statement – a proposition expressed in language – that applies, at least potentially, to more than one case. Usually, this is marked by the occurrence of expressions with appropriate quantifiers, such as "for all X, subject to certain conditions, Y is the case". A principle is *evaluative* if it has evaluative content; for instance, it includes evaluative predicates or concepts such as *good* or *bad*, *better* or *worse*, *desirable* or *undesirable*. A principle is *normative* if it has normative content; for instance, it includes deontic operators such as *ought*, *may*, *permissible*, *obligatory*, *right*, or *wrong*. A principle without any evaluative or normative content is *positive*. Such principles are common in the sciences; think of the principle of conservation of energy in physics.

Classic examples of normative principles are the Ten Commandments from the Bible, the Golden Rule ("You should treat other people in the way in which you would like them to treat you"), Bentham's principle of utility ("An action is right if it maximizes total utility"), and more recently Rawls's principles of justice. Roughly speaking, these state that each person is entitled to the most extensive system of individual liberties, compatible with a similar system for everyone else (the "equal-liberty principle"); and socio-economic inequalities are permissible if and only if they are compatible with a system of fair equal opportunities (the "fair equality of opportunity principle") and benefit the least well-off members of society (the "difference principle") (Rawls 1971/1999).

The *propositional content* of a principle is the set of all its implications. Sometimes this propositional content may depend on auxiliary assumptions. As already noted, their having propositional content sets principles apart from concepts, which merely offer categorizations.

3.2 What is a theory?

The word "theory" is commonly used in two distinct senses. First, it can refer to an entire academic field or area of enquiry, such as when we speak of "political theory" or "economic theory" as general areas to which curricula or scholarly journals are devoted. Second, the word can refer to a specific theory within such an area, such as Rawls's theory of justice, the theory of the firm in economics, or Newton's theory of physics. Our focus here is on theories in this second, specific sense. Surprisingly, there exists no canonical definition of a theory in that sense in political theory.

To provide a starting point, we propose a simple definition inspired by the philosophy of science but adapted to the present context.[8] We define a *theory* as a set of statements – propositions expressed in language – which is a candidate for playing some theoretical or practical role and which is, ideally, representable as the set of all implications of some underlying principles. The set of principles from which the theory can be derived – if there is such a set – is called the *theory formulation*.

Although loose and abstract, this definition has some merits. First, it allows us to view positive theories (in the sciences) and normative or evaluative theories (in moral philosophy or political theory) as instances of the same general category. Second, it makes transparent the differences between them. For example, the roles played by theories can range from descriptive, explanatory, and predictive (in the case of positive theories) to evaluative and prescriptive (in the case of evaluative or normative theories). Third, the definition allows us to identify the special challenges that arise when we construct and assess normative or evaluative theories.

Paradigmatic examples of theories according to our definition are Newton's theory of physics and Rawls's theory of justice. Each can be viewed as a set of statements, entailed by some underlying principles, which can play a descriptive, explanatory, or prescriptive role. Newton's theory is the set of all statements entailed by Newton's principles of physics, perhaps together with some empirical premises about the solar system or other physical systems of interest. It can be used, for instance, to explain and predict the trajectory of the planets around the sun and to guide such engineering projects as travelling to the moon and safely back. Rawls's theory is the set of all statements entailed by Rawls's principles of justice, perhaps together with some empirical premises about relevant social conditions. It can play a prescriptive or normative role, guiding us in the design of social institutions (for an earlier discussion of the relationship between normative and positive theories, see McDermott 2008).

For our purposes, the biggest structural difference between Newton's and Rawls's theories is that one is positive and the other normative. Indeed, the principles underlying Newton's theory are positive principles, while those underlying Rawls's are normative ones. Generally, a theory is *positive* if it has no evaluative or normative

---

[8] We here follow broadly what is often called the *syntactic* approach to defining theories (where a theory is defined as a set of sentences/propositions with certain properties); it is arguably the most conventional approach. For a classic exposition, see Quine (1975). It is also worth exploring the rival *semantic* approach (where a theory is defined as a set of models with certain properties), but given space constraints, we set this aside here (see van Fraassen 1980). For an introduction to the philosophy of science, see Okasha (2002).

content; it is *evaluative* or *normative* if it has such content. Evaluative theories that offer evaluations of "goodness" or "betterness" are also called *axiological*.

Earlier, we associated evaluative content with the occurrence of evaluative predicates or concepts, such as *good*, *better*, *desirable*, and so on, and normative content with the occurrence of deontic operators, such as *ought*, *may*, and so on. Different accounts of what qualifies as evaluative or normative content can be given, and we need not commit ourselves to one such account here. On any reasonable account, Rawls's theory will come out as normative and Newton's theory as positive.

While Newton's and Rawls's theories are paradigmatic instances of our definition, a theory need not be self-consciously theoretical. A set of rough and informal principles describing how ordinary objects behave when pushed, dropped, or thrown can constitute a "folk" theory of motion that is predictively useful in everyday contexts. Similarly, a set of basic principles describing how animals respond to noise, movement, and the presence of humans may constitute a simple predictive theory of animal behaviour that members of hunter-gatherer societies might have used to guide their actions. We also routinely employ normative theories without self-consciously doing so. For example, a set of simple principles specifying how we should or should not treat others may constitute a simple "folk" theory of personal ethics.

3.3 The distinction between a theory and the support for it

Our simple definition of a theory is silent on whether the theory is true, useful, or good in some sense. A theory that is false, irrelevant, or superseded still counts as a theory. Describing something as a theory carries no assessment of its truth or acceptability. Thus the most far-fetched and implausible conspiracy theory – for example, in science or in history – still qualifies as a theory.

Similarly, the attempt to downgrade a set of statements by asserting "it is *only* a theory", as critics of evolutionary theory or global-warming sceptics sometimes do when they describe their target, makes little sense. Calling something a theory is consistent with its being well-supported and true just as it is consistent with its being speculative and even false.

An important distinction is that between a theory and what is offered in its support. A theory, as we have defined it, is distinct from any arguments, evidence, or justification given for it. For example, Newton's actual or hypothetical experiments – such as how an apple fell onto his head (though supposedly a myth) – are not part of his physical theory itself; they are part of the evidence he had for that theory. Similarly, Rawls's arguments in support of his principles of justice, such as his original-position thought experiment (a hypothetical choice situation in which the parties to the social contract must agree on principles governing the basic structure of society), are not strictly speaking part of his theory of justice itself. The original-position thought experiment, like any physical experiment or any social scientist's observation, is offered in support of the theory in question. Thus Rawls's book, *A Theory of Justice*, consists of the theory itself (everything entailed by his principles of justice), together with arguments in support of it (most notably, the original-position thought experiment and reflective-equilibrium considerations), comparisons with rival theories (such as utilitarianism), and a fair amount of commentary on the theory's interpretation and its

applications. This distinction, between a theory and the argument or evidence in support of it, will help us clarify some recent methodological debates in political theory, about abstraction and idealization, which we discuss in Section 5 below.

An important task, separate from defining a theory, is therefore to identify the requirements on a good, or acceptable, theory, and to spell out how we can assess it.

4. The assessment of principles and theories

There are two kinds of criteria that we may use to assess – especially to justify or to criticize – principles and theories: "internal" and "external" criteria. The former concern the way the principles or theory are formulated and their internal logical structure. Criteria such as consistency and parsimony fall into this category. The latter concern the relationship between the principles or theory and what these are "about": their normative or evaluative content, in analogy with the empirical or descriptive content of a scientific theory. Criteria such as truth or normative adequacy (in analogy with truth or empirical adequacy in science) fall into this second category. In this section, we discuss the two kinds of criteria ("internal" and "external") in turn.

Throughout this discussion, we focus on theories, rather than principles, as the units of assessment. This is no loss of generality. We are usually interested, not in individual principles in isolation, but in sets of principles that we wish to assess together. If we wish to assess a principle by itself, we can view it as a special case of a theory, namely a theory that consists just of the principle and its implications.

4.1 Internal criteria

Although we have kept our definition of a theory deliberately thin, defining it simply as a set of statements that may play some theoretical or practical role and that is, ideally, derivable from some underlying principles, we usually want theories to satisfy some further requirements. We now discuss several common criteria for assessing a theory's internal structure. Like our definition of a theory, they are inspired by the sciences, but apply to normative and evaluative theories as much as they apply to positive ones. Some of the criteria are so obvious that they are often left unacknowledged; nonetheless, it is useful to make them explicit.

Consistency: We require a good theory to be *logically consistent*. Formally, the set of statements constituting the theory must be capable of being simultaneously true. An obvious reason for requiring consistency is that anything follows from an inconsistent set of statements (*ex falso quodlibet*). Thus an inconsistent set, such as one containing both "*p*" and "*not p*", is of little use, whether for explanatory, predictive, evaluative, or prescriptive purposes. By entailing everything, it is too indiscriminate.[9]

Deductive closure: We require a good theory to be *deductively closed*. This means that any statement that is logically entailed by the theory also belongs to the theory.[10]

---

[9] Sometimes we may be prepared to lower the bar of consistency, by admitting some "contained" or "local" inconsistencies, as in so-called "paraconsistent" logics. Even then, we usually impose some weakened variant of a consistency requirement, to rule out those inconsistencies that are too global to preserve a theory's usefulness.

[10] A set of statements is *deductively closed* if it contains all its implications.

The idea underlying deductive closure is that we want to be able to identify a theory with everything the theory is logically committed to. If the theory asserts "*p*" and "*if p then q*", for example, then it should also assert "*q*". We would consider a theory defective if it were committed to the first two statements, but not to the third. The way we characterized Newton's and Rawls's theories had deductive closure built into it, since we characterized each as the set of all statements that are logically entailed by the relevant principles. Deductive closure is easy to achieve even when a theory is initially given in a form that violates it: we can re-define the theory as the set of all statements entailed by the original, non-deductively-closed formulation.[11]

Axiomatizability: This requirement is implicit in the final clause of our definition of a theory, which says that a theory should ideally be representable as the set of all implications of some underlying principles. Formally, a theory is *axiomatizable* if there exists a finite set of principles such that the entire theory can be expressed as their body of implications.[12] It should be evident from our earlier discussion that Newton's and Rawls's theories are usually presented in axiomatized form, as the bodies of implications of Newton's and Rawls's principles, respectively, perhaps together with some auxiliary assumptions. An axiomatizable theory can be presented in an informative manner, simply by specifying the set of principles from which it can be derived. If the theory could only be presented by brute enumeration of all its implications – typically infinitely many – there would be no succinct way of summarizing its content. A good theory illuminates its subject matter by giving us a manageable set of principles – a manageable theory formulation – that encodes the theory's entire propositional content.

Parsimony: We require a good theory to avoid any unnecessary complexity, and to be as simple as possible, in an appropriate sense of simplicity (on simplicity, see, e.g., Baker 2013).[13] Scientists commonly care about parsimony, often under the label "Occam's razor", and Rawls, for instance, also emphasizes simplicity as a virtue of a normative theory.[14] What counts as "simple" may be different from context to context and may also depend on what the theory is about. We usually want to find the simplest theory able to account for its subject matter. To be illuminating, the theory ought to be simpler, for example, than the target phenomenon it seeks to account for, as well as simpler than its rival theories. A scientific theory, for instance, should be simpler and more succinct than an enumeration of all the empirical facts it seeks to explain; otherwise it cannot play any explanatory role. Likewise, a good normative theory should be simpler than an enumeration of all case-specific normative

---

[11] The deductive-closure requirement also highlights, once more, why consistency matters. Since an inconsistent set of statements *entails* everything, deductive closure would force such a theory to *consist* of everything that can be expressed in the relevant language, which would amount to a completely uninformative theory. Again, if we were to use a paraconsistent logic, we might weaken the requirement of deductive closure but still retain some less demanding requirement in a similar spirit.

[12] See also Quine (1975). Note that any axiomatizable theory is deductively closed. Note, further, that logicians sometimes replace the "finiteness" requirement with a weaker "formal decidability" requirement; we set these technicalities aside.

[13] Although axiomatizability of a theory in terms of some easily expressible principles is one of the marks of parsimony, axiomatizability is not a sufficient condition for parsimony.

[14] In his discussion of why his principles of justice are preferable to utilitarian principles, Rawls says that "reasonable risk aversion may be so great, once the enormous hazards of the decision in the original position are fully appreciated, that the utilitarian weighting may be, for practical purposes, so close to the difference principle as to make the *simplicity* of the latter ... decisive in its favour" (Rawls 1999, 144 emphasis added).

judgments. The relevant bar of simplicity may be adjusted depending on the theory's subject matter.

The present list of criteria for the internal assessment of theories is only illustrative, not exhaustive, but given space constraints, we now move on to external criteria.

4.2 External criteria

We defined a theory as a set of statements that is a candidate for playing some theoretical or practical role. Implicit in this definition is the idea that there is something the theory is about: any theory is intended to represent, summarize, or capture something "outside the theory". It may capture this correctly, in which case the theory is true, correct, or externally valid, or it may fail to do so, in which case it is false, incorrect, or externally invalid.

What exactly a given theory is intended to represent needs to be spelt out further. In the case of a physical theory, the answer is relatively straightforward, especially if we accept "scientific realism": it is intended to represent certain physical facts about the world, such as facts about how physical objects behave in response to each other (see, e.g., Chakravartty 2013). In the case of a normative or evaluative theory, the picture is more complicated. If we are realists about normative or evaluative matters, we may say that the theory is intended to represent some theory-independent moral facts. If we are not realists about normative or evaluative matters, it is harder to specify what a normative or evaluative theory is intended to represent (on moral realism and anti-realism see, respectively, Sayre-McCord 2011; Joyce 2009).

Yet the very idea of a theory breaks down unless we assume that there is something potentially representable by it, however observer-dependent or socially constructed it might be.[15] If we were nihilists, to take an extreme example of the denial of any normative or evaluative facts, we would not be able to engage in normative or evaluative theorizing in earnest.

Thus, in this section, we assume that normative or evaluative theories are *truth-apt*: it makes sense to ask whether they are true – or, some might prefer to say: correct or externally valid. We thus accept a form of "cognitivism" about such theories. This assumption is still compatible with a variety of views about the "meta-ethical" status of normative or evaluative judgments. We return to some of these issues in Section 6. Granting, then, that there is some standard of correctness by which we can assess normative or evaluative theories (an ontological assumption), we still need to know how to do this assessment (an epistemological question). We now review several methods of testing a theory for external validity.

Taking intuitive judgments as strict evidence: According to this method, the test for a normative or evaluative theory is whether it fits our intuitive judgments about the relevant normative or evaluative matters. On this approach, our intuitive judgments

---

[15] A constructivist might take a theory to represent certain constructed facts. This is consistent even with the view that the theory itself is the "vehicle" by which those facts are being constructed. To develop that view further, one might draw, for instance, on parallels with Searle's analysis of declarative speech acts (which – roughly – bring certain facts into existence by representing them). See Searle (1995).

have the same status as empirical observations in science. In science, a theory is *empirically adequate* if it entails the correct observation statements (see, e.g., Quine 1975; van Fraassen 1980). Similarly, in moral and political theory, we might call a theory *normatively* or *evaluatively adequate* if it entails the correct normative or evaluative statements. According to the strict-evidence method, these are precisely the normative or evaluative statements supported by our intuitive judgments (cf. the discussion in Dworkin 1975). Although simple and analogous to familiar scientific methods, this method has some problems. First, while we may be confident in some of our normative or evaluative judgments, other judgments may be more tentative, and in some cases – especially when the issue is less familiar – we may not have any firm intuitions at all. Second, our intuitive judgments may be subject to biases and framing effects, which may cast further doubt on their reliability. Third, our normative or evaluative judgments may not be consistent with one another, or they may entail other judgments that we reject on reflection; in such cases, the strict-evidence method provides no guidance at all.

Reflective equilibrium: A more refined method of theory testing is the reflective-equilibrium method. It does not treat our intuitions as independent evidence, prior to the theory in question, but requires us to reach a "mutual fit" between the theory and our considered judgments. This works as follows. We begin with some initial theory, perhaps inspired by our initial intuitive judgments or given by some *prima facie* principles, then consider the implications of the theory and ask whether they are also in line with our judgments. If those implications fit our judgments, the process stops. It is more likely, however, that only some of the theory's implications fit our judgments, while others do not. We then reassess both the theory and our judgments. In some cases, we may decide, on reflection, to revise the theory by changing some of the constituent principles, so as to bring the theory in line with the judgments we are unwilling to give up. In other cases, we may decide to overrule our judgments and embrace the theory's implications as our new considered judgments. A *reflective equilibrium* is reached when the implications of our possibly revised theory are in line with our possibly revised judgments. At least since Rawls's *Theory of Justice*, reflective equilibrium has been one of the most widely used methods in political theory (Rawls 1999, 15–18, 40–46; Daniels 2013). Its details can be spelt out in a variety of ways. For example, we may choose the units of assessment more narrowly or more broadly. We can either search for a reflective equilibrium involving a narrowly specified theory, constituted by a small number of principles together with very few auxiliary assumptions, or we can search for a reflective equilibrium involving a more broadly specified theory, constituted by a larger number of principles and further additional assumptions – which in turn may be interpreted as a conjunction of multiple theories, covering a wider domain of issues. Political theorists sometimes speak of "narrow" reflective equilibrium in the first case, and "wide" reflective equilibrium in the second (Daniels 2013). Similarly, we may take different views on which kinds of judgments – especially whose judgments – should serve as input to this method; more on this below. Although the reflective-equilibrium method is consistent with the idea that we arrive at our normative or evaluative theories through careful deliberation, we may be worried about the possible arbitrariness of its outcome, since there may not always be a unique equilibrium. In some cases, we may not be able to reach any equilibrium at all (the *non-existence* problem), such as when we theorize about genuine moral dilemmas and vacillate between different theories that each fit only some of our judgments while conflicting with others. In other cases,

there may exist more than one equilibrium, in that we can arrive at different "packages" of revised theories and judgments that each have the required "mutual fit" (the *non-uniqueness* problem). Arbitrary factors such as framing effects or the order in which we consider different implications of the theory may then affect which reflective equilibrium we end up with (the *path-dependence* problem).

Thought experiments and intuition pumps: Whether we opt for the strict-evidence method or the reflective-equilibrium method, we may sometimes wish to sharpen or clarify our intuitions or judgments. Thought experiments and real-world cases can serve as useful "intuition pumps" (Dennett 2013; Brownlee and Stemplowska forthcoming). Here, we consider some hypothetical or actual scenario that prompts strong normative or evaluative judgments. In the much-discussed "trolley problems", for example, we are asked to judge what actions, if any, would be permitted to prevent a run-away trolley from crashing into, and killing, a larger group of people, at the expense of leading it to crash into, and kill, a smaller group (Thomson 1985). We then use these judgments to test our relevant normative or evaluative theories, following either the strict-evidence method or the reflective-equilibrium method. The usefulness of intuition pumps, especially ones involving highly idealised, counterfactual scenarios has recently been the object of considerable controversy in political theory (Elster 2011). We return to this issue in Section 5.

The relevant judgments: Both the strict-evidence method and the reflective-equilibrium method raise the question of which kinds of judgments, and whose judgments, to use in testing our theories. Should we test our theories on the basis of relatively spontaneous judgments or on the basis of suitably "filtered" judgments, and how should that filtering take place (Rawls 1999, 42)? And should we use the political theorist's judgments (which might be affected by his or her ideological views) or society's (which might similarly be affected by biases), and in the latter case, which society should we focus on (Miller 1992; Walzer 1983)? For example, while the Rawls of *A Theory of Justice* arguably followed the former approach (relying on the political theorist's judgments), the Rawls of *Political Liberalism* subscribed to the latter (looking at society) (Rawls 1971/1999; Rawls 1996). Specifically, the later Rawls re-interprets his theory of justice as an articulation of the ideas implicit in the public culture of liberal democratic societies. The building blocks of his account of justice are explicitly "drawn" from, and supposed to be widely acknowledged within, the society for which that account is designed. Relatedly, political theorists disagree about whether the judgments they appeal to in theory testing should "fit" the particular practice the theory is meant to regulate. If the answer to this question is positive, then the exercise of theory construction is best seen as an attempt to offer what Ronald Dworkin calls a "constructive interpretation" of existing political practices (Dworkin 1986, chap. 2; James 2005; Sangiovanni 2008). If the answer is negative, then the exercise of theory construction is best understood as an attempt to "discover" particular normative and evaluative truths, independently of existing social practices (on interpretation versus invention, see Walzer 1987). To illustrate the difference between the two approaches – at least on one reading of what sets them apart – consider the following judgment, made explicitly in relation to the practice of camping with friends: "we should institute a regime of shared ownership" (see G. A. Cohen 2009). On a practice-dependent/interpretive approach, this judgment should count as relevant evidence only in the construction of a "normative theory of camping", not in the construction of

normative theories of other practices, such as socio-political relations within the state. For proponents of practice-independence, by contrast, all normative or evaluative judgments, including the one in question, have cross-contextual validity in the identification of what justice or other moral values demand (this point is made in Ronzoni 2012; see also Miller 2002).[16]

The applied-moral-philosophy method: A final method of justifying a theory in political theory is to show that it can be derived from some independently accepted moral principles or theory. A committed utilitarian or Kantian, for example, may regard a normative or evaluative theory as justified if and only if it can be derived from utilitarian or Kantian principles, which are treated as independently given. In this vein, Nozick says: "Moral philosophy sets the background for, and boundaries of, political philosophy. What persons may and may not do to one another limits what they may do through the apparatus of the state, or do to establish such an apparatus" (Nozick 1974, 6; see also Otsuka 2003, 3). Those theorists who view political theory as a subfield of moral philosophy will find this method appropriate. (Works in which political theory is conducted – at least to some extent – as applied moral philosophy include Singer 1972; G. A. Cohen 2008; Fabre 2012.) By contrast, if we consider the activity of political theory to take place against the background of (reasonable) pluralism about moral matters, the applied-moral-philosophy method is problematic, since it relies on the acceptance of a specific moral theory (for a critique of the "applied moral philosophy" approach, see Williams 2005; see also Galston 2010).

5. Abstraction and idealization in political theory

In political theory, as well as in other disciplines, theories are often abstract and/or idealised in certain respects. In this section, we explain what this means and discuss some methodological issues raised by abstraction and idealization.

5.1 Defining abstraction and idealization

Broadly following Onora O'Neill (1996, chap. 2), we say that a theory is *abstract* with respect to an issue – represented by a set of statements – if it is silent on that issue; formally, it has no implications at all for the given statements, implying neither any of these statements nor any of their negations. Newton's theory of physics, for example, is abstract with respect to the colours of the physical bodies whose motion it represents. A theory is *idealized* with respect to an issue – again represented by a set of statements – if it entails some (simplifying or limiting) falsehood about that issue; formally, it has a false implication for some of the given statements, implying one or more false statements among them or the negations of one or more true statements. Newton's theory of physics, for instance, is idealized with respect to friction, entailing the absence of friction in the physical systems it represents, although friction is present in the real world.

Since theories are meant to simplify the world, most political theorists agree that abstraction is virtually unavoidable in theory construction, and an innocuous intellectual exercise (O'Neill 1996, chap. 2; see also the discussion in Stemplowska

---

[16] The distinction between practice-dependence and practice-independence is complex and much debated. Due to space constraints, we are unable to explore this complexity here.

2008). Idealization, unlike abstraction, is looked at with greater suspicion in political theory and is considered potentially problematic (O'Neill 1996, chap. 2).

5.2 The worry about idealization

The "danger of idealization" in political theory has been discussed primarily in the debate on "ideal versus non-ideal theory" (for an overview, see Valentini 2012). The debate is largely animated by the worry that resort to simplifying assumptions and idealized thought experiments or intuition pumps – which are common in contemporary political theory – will adversely affect the validity of the ensuing theories (see, e.g., O'Neill 1996, chap. 2; Farrelly 2007; Mills 2005). For instance, John Rawls develops his theory of justice assuming full compliance with the resulting principles. This makes critics wonder whether we can trust Rawls's principles to deliver correct and action-guiding prescriptions for the real world, where many people fail to abide by the demands of justice.

In assessing this worry, two points are worth making. First, like abstraction, some degree of idealization in our theories – in the form of simplifying assumptions – may play an important, and justified, heuristic role. Again, Newton's theory of physics does not appear to be significantly undermined by its assuming friction away, especially to the extent that information about friction can potentially be re-introduced in refinements or applications of the theory. This suggests that, rather than reject idealizations as problematic from the outset, we must ask whether a theory contains the "right" idealizations, given its purpose (Robeyns 2008; Valentini 2009).

Second, to answer the question of which idealizations are "right" and which are not, we need to distinguish between three possible *loci* – or levels – at which idealizations can occur: (1) the theory itself, (2) the conditions of application of the theory's prescriptions (where those prescriptions are of the form "if such-and-such conditions hold, then such-and-such follows"), (3) the justification of the theory. Crucially, idealization at any one of these levels need not entail idealization at any of the others.

5.3 Rawls's theory as an example

Rawls's theory of justice, as described earlier, consists of the "equal liberty" (X), "fair equality of opportunity" (Y), and "difference" (Z) principles and their implications. The theory itself would be idealized – a level (1) idealization – *if and only if* these principles entailed false statements about their subject matter – for instance, if the statement "we ought to re-arrange the tax system so as to benefit the worst off as much as possible" were false. Note, however, that the often-criticized "idealized" assumption of full compliance is not made within the theory itself, but occurs as part of the *justification* that Rawls offers for the theory. It is one of the assumptions made by the parties in the original-position thought experiment. So, it is an idealization at level (3). Rawls's principles, which generate his theory, do not imply any false claims about full compliance; hence we have no idealization at level (1) here. Nor do the prescriptions following from these principles presuppose full compliance for their *applicability*; hence we have no idealization at level (2) either (see Simmons 2010, 9–10 for discussion).

Similarly, consider Rawls's assumption that society exists under favourable historical and social conditions. Relative to existing war-torn, or desperately poor countries, this assumption is clearly false: it is an idealization. Rawls says explicitly that his principles of justice may not apply to societies in which the relevant favourable conditions are absent (Rawls 1999, 216). Does this make Rawls's *theory* idealized – an idealization at level (1) – and problematically so? Arguably, it does not, because the favourable-conditions idealization operates at the level of the *conditions of application* of the theory's prescriptions, i.e., at level (2). The prescription that is entailed by the theory (under a careful formulation) – namely "*if favourable conditions hold*, justice demands X, Y, and Z" – is still true. Although the conditional nature of this prescription limits the scope of application of the theory, it does not make the theory itself idealized, by generating *false* prescriptions.

5.4 The worry reassessed

We suspect that, although worries about idealization in political theory are frequently expressed as complaints about "*theories* being idealized", they actually tend to target idealizations at levels (2) and (3), rather than (1) – i.e., at the levels of the theories' conditions of application and the justifications offered for those theories, not at the level of the theories themselves.

Of course, idealizations at levels (2) and (3) often make the target-theories somewhat irrelevant to the real world, by rendering them insufficiently action-guiding in real-world circumstances. While this lack of guidance may be a genuine shortcoming, it is not accurately captured by the claim that the theories themselves are idealized. Rather, in the case of a level (2) idealization, the theories entail true prescriptions of an "if-then" sort, whose antecedent conditions – the "if" clauses – do not hold in real-world circumstances. And in the case of a level (3) idealization, the theories may be insufficiently justified, in that the justifications offered for them – such as highly contrived thought experiments – are too idealized to allow inferences for the real world.

Our discussion highlights the importance of clarity about the role that idealizations play in the defence and formulation of one's theory. Consider a theory of justice prescribing "*p*", defended on the assumption that there is no reasonable disagreement about justice within society, and yet formulated in universal terms: "justice *always* demands that *p*". This theory does indeed run the risk of being problematically idealized, if it turns out that the presence of reasonable disagreement makes a morally relevant difference to what justice demands. The theory, in that case, would have false implications in a number of situations, namely those involving reasonable disagreement. A true principle, by contrast, would only say: "if there is no reasonable disagreement, justice demands that *p*".

6. The significance of disagreement in political theory

We noted earlier that, while political theory is sometimes viewed as a subfield of moral philosophy, another view is that political theory is distinct from moral philosophy in that the conditions of theorizing are different (see, e.g., Williams 2005; Larmore 2013). On this view, political theory, unlike – or much more than – moral philosophy, is conducted against the background of (reasonable) pluralism in society.

Disagreement about normative and evaluative matters, it is said, has a different status in political theory than in (mainstream) moral philosophy. The nature of disagreement in society constrains what normative or evaluative principles can be defended in political theory. In this section, we explore this view and discuss its methodological implications.

6.1 Disagreement: "constitutive of assertability conditions" rather than "epistemic"

While moral philosophers seek to answer questions such as "what ought we to do in a given situation", political theorists are faced with questions such as "what ought we do, given that we do not agree about what we ought to do" (see, e.g., Waldron 1999; Larmore 2013). At first sight, we might be perplexed by this alleged difference between political theory and moral philosophy. Are there not equally big disagreements in moral philosophy? Practically every well-known moral dilemma has the property that different people take different views on how it should be resolved, and the question of what one should do in such cases of disagreement arises also in the moral realm.

Yet, the difference between moral philosophy and political theory may be said to be the following. In moral philosophy, we commonly (though not universally) make the assumption that, among the many different rival normative or evaluative theories, one is the independently correct or true theory. The task for the moral philosopher is to identify that theory. *If* we conduct moral philosophy on this assumption, disagreement is of a "merely" epistemic kind. There is a fact about what the right answer to any normative or evaluative question is; we may just have different beliefs about that fact.

It is less clear – so the argument goes – whether the same assumption can be made in political theory. On this picture, disagreement in political theory may be viewed, not merely as "epistemic", reflecting different beliefs about the same truth, but as partly "constitutive" of the correctness conditions – or warranted assertability conditions – of normative claims themselves. On this view, whether a normative or evaluative theory in political theory is correct (or assertable) depends, in part, on the society in which the theory is to be applied, and specifically on the level of reasonable disagreement in that society; we define the notion of *reasonableness* below.

A key desideratum here is that our theories be, at least in principle, acceptable to individuals holding reasonable but conflicting moral views within pluralistic societies. What counts as a correct normative or evaluative principle in a society with little reasonable disagreement need not always count as correct in a highly pluralistic society, where the range of reasonable disagreement is greater. In light of this, some theorists avoid using the notion of truth in political theory altogether, and prefer to replace it with other, less loaded notions, like reasonableness or reasonable acceptability (Rawls 1996; cf. J. Cohen 2009).

But what might justify the shift from the "epistemic" to the "constitutive" status of disagreement in the political domain? At least two answers are possible. First, the shift might be *morally justified*. On this view, a commitment to respect for persons places a burden of "reasonable acceptability" on the principles put forward by the political theorist, insofar as these principles may be *permissibly enforced* against individuals, for instance through state action. The idea is that political theory focuses

on the development of "enforceable rules", and these rules are normatively appropriate – meet the relevant criterion of correctness – only if they are acceptable to reasonable individuals to whom they apply.

Second, reasonable disagreement may be taken to constrain correctness or warranted assertability in political theory for *pragmatic reasons*. After all, most of the principles put forward by political theorists are meant to help regulate social life in complex and highly pluralistic societies. A normative or evaluative theory that appeals only to people of a particular moral persuasion would be of little use in this respect; it could not give rise to a durable and stable social order except through autocratic imposition. And so again, acceptability to individuals with competing reasonable views may be deemed a criterion of correctness of the theory.

The inspiration for many of the foregoing reflections can be found in John Rawls's second major book, *Political Liberalism* (1993/1996). We now offer a brief discussion of Rawls's treatment of the relevance of pluralism to political theorizing.

6.2 The distinction between "political" and "comprehensive" theories

One of the key innovations of *Political Liberalism*, compared to *A Theory of Justice*, is Rawls's insistence on his theory being "political not metaphysical" (Rawls 1985). This means two things. First, the theory does not adjudicate all aspects of interpersonal conduct (as a "comprehensive" moral theory might do) but concerns only the "public" realm (most notably, the basic structure of society). Second, the theory is grounded *not* in any "comprehensive" moral theory ("doctrine"), on which there is likely to be deep disagreement in any pluralistic liberal society, but in ideals that are drawn from the public culture of such a society. Thus a "political" theory is typically (i) restricted to a smaller domain of issues and (ii) less morally and metaphysically "loaded" (thereby "thinner") than a "comprehensive" theory. In particular, it refrains from taking a stand on issues that are too controversial.

For example, a theory of justice that adjudicates all aspects of personal, and not just political and social, life would be "comprehensive", as would be a theory based on the metaphysically loaded premise that human beings are morally equal because they are created "in the image of God". Both theories would be the object of *reasonable disagreement* within a pluralistic society. By contrast, a theory of justice that focuses on the basic structure of society and is based on a commitment to "citizens' freedom and equality" would count as "political", since its domain does not "over-reach" and its premises are entrenched in the public culture of liberal democracies and arguably shared by individuals holding competing yet reasonable comprehensive doctrines. But what does the notion of "reasonableness" stand for?

6.3 The notion of "reasonable disagreement"

The notion of "reasonableness" is hard to pin down in general, and this is true in Rawls's case as well (for discussion, see Gaus 1996, 131–132; Gaus 1999). In particular, the notion may be interpreted in epistemological and/or moral ways. Under an epistemological interpretation, something is *reasonable* if it is consistent with a proper use of reason in light of the evidence available. Rawls believes that, because of what he calls the "burdens of judgment", we should not expect agreement between the

views of different people, even when they are all developed by consistently applying the powers of reason (Rawls 1996, Lecture 2, sec. 2). Under a moral interpretation, *reasonableness* refers to a view's compatibility with certain fundamental normative requirements, such as respect for citizens as free and equal, or a commitment to mutual justification (Rawls 1996, Lecture 2, sec. 3). This would make views that reject such a commitment unreasonable.

Both interpretations of "reasonableness" have virtues and vices, especially in the context of what we earlier called the "moral" justification for deeming acceptability in the context of reasonable disagreement an assertability condition for theories within political theory. As far as the epistemological interpretation of reasonableness is concerned, on the positive side, it offers a compelling rationale for treating only *some* contested views as "worthy of respect", namely those with a given epistemic pedigree. On the negative side, this interpretation potentially allows morally repugnant views to count as reasonable, assuming they involve no breach of reason, purely epistemically understood. If a necessary condition for the overall acceptability of a normative or evaluative theory is that it is acceptable to all individuals with reasonable views, an acceptable theory must then potentially appeal even to individuals with morally repugnant views, and this may be too much to ask.

As far as the moral interpretation of reasonableness is concerned, on the positive side, it allows us to "filter out" morally repugnant views. But, on the negative side, it does so at the cost of referring to substantive values that themselves require justification. The suspicion is that, on a moralized interpretation, the "reasonable" simply corresponds to the moral commitments that the liberal theorist considers non-negotiable: it is *ad hoc* (see Mouffe 2005). Again, we do not take a stand on this, but note that these difficulties threaten the plausibility of defining the overall acceptability of a normative or evaluative theory simply in terms of its acceptability to all individuals with reasonable views; identifying those reasonable views may itself rely on moral premises.

6.4 The quest for an "overlapping consensus"

Finally, turning to the "pragmatic" justification for taking disagreement to have a "constitutive" rather than purely "epistemic" status in political theory, we conclude with some remarks on Rawls's notion of an "overlapping consensus". An *overlapping consensus* on a normative or evaluative theory in a social or political domain occurs when this theory is endorsed from the perspective of different reasonable comprehensive doctrines held in society. A normative theory that cannot be endorsed from different such perspectives could not hope to gain sufficient support and to offer a stable basis for social organization in pluralistic liberal democracies (Rawls 1996, Lecture 4). The notion of an overlapping consensus operationalizes the idea that the criterion of correctness of a "political" theory is its acceptability to individuals who hold competing but reasonable comprehensive moral theories.

7. Concluding remarks

We have reviewed the methodology of analytic political theory from what we hope is a somewhat novel and helpful angle. By drawing on ideas from the philosophy of science, we have attempted to highlight the ways in which theorizing in political

theory relates to theorizing in other areas of philosophy and positive science. We have also reviewed some recent debates and controversies within the political-theory literature, which has only recently given greater attention to methodological questions. Our hope is that this article will prove to be a clarifying contribution to the growing methodological debate in political theory.

References

Arrow, Kenneth Joseph. 1951. *Social Choice and Individual Values*. New York: John Wiley & Sons.

Baehr, Amy R. 2013. "Liberal Feminism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2013. http://plato.stanford.edu/archives/win2013/entries/feminism-liberal/.

Baker, Alan. 2013. "Simplicity." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2013. http://plato.stanford.edu/archives/fall2013/entries/simplicity/.

Blackorby, Charles, Walter Bossert, and David J. Donaldson. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. New York: Cambridge University Press.

Bratman, Michael. 1999. *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press.

Brownlee, Kimberley, and Zofia Stemplowska. forthcoming. "Trapped in an Experience Machine with a Famous Violinist: Thought Experiments in Normative Theory." *N/A*

Carter, Ian. 2012. "Positive and Negative Liberty." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2012. http://plato.stanford.edu/archives/spr2012/entries/liberty-positive-negative/.

Chakravartty, Anjan. 2013. "Scientific Realism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2013. http://plato.stanford.edu/archives/sum2013/entries/scientific-realism/.

Christiano, Thomas. 2008. "Democracy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2008. http://plato.stanford.edu/archives/fall2008/entries/democracy/.

Cohen, G. A. 1988. *History, Labour, and Freedom: Themes from Marx*. Oxford: Clarendon Press.

———. 1995. *Self-Ownership, Freedom, and Equality*. Cambridge: Cambridge University Press.

———. 2008. *Rescuing Justice and Equality*. Cambridge, MA: Harvard University Press.

———. 2009. *Why Not Socialism?* Princeton, N.J.: Princeton University Press.

Cohen, Joshua. 2009. "Truth and Public Reason." *Philosophy & Public Affairs* 37 (1): 2–42.

Daniels, Norman. 2013. "Reflective Equilibrium." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2013. http://plato.stanford.edu/archives/win2013/entries/reflective-equilibrium/.

Dennett, Daniel C. 2013. *Intuition Pumps And Other Tools for Thinking*. New York: W. W. Norton.

Dworkin, Ronald. 1975. "The Original Position." In *Reading Rawls: Critical Studies on Rawls's A Theory of Justice*, edited by Norman Daniels, 16–52. Stanford, CA: Stanford University Press.

———. 1986. *Law's Empire*. Cambridge, MA: Harvard University Press.

Elster, Jakob. 2011. "How Outlandish Can Imaginary Cases Be?" *Journal of Applied Philosophy* 28 (3): 241–258.

Fabre, Cécile. 2012. *Cosmopolitan War*. Oxford: Oxford University Press.

Farrelly, Colin. 2007. "Justice in Ideal Theory: A Refutation." *Political Studies* 55 (4): 844–864.

Gallie, W. B. 1955. "Essentially Contested Concepts." *Proceedings of the Aristotelian Society* 56 (1): 167–198.

Galston, William A. 2010. "Realism in Political Theory." *European Journal of Political Theory* 9 (4): 385–411.

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Gaus, Gerald F. 1996. *Justificatory Liberalism : An Essay on Epistemology and Political Theory*. Oxford: Oxford University Press.

———. 1999. "Reasonable Pluralism and the Domain of the Political: How the Weaknesses of John Rawls's Political Liberalism Can Be Overcome by a Justificatory Liberalism." *Inquiry* 42 (2): 259–284.

Gerring, John. 1999. "What Makes a Concept Good? A Criterian Framework for Understanding Concept Formation in the Social Sciences." *Polity* 31 (3): 357–393.

Gilabert, Pablo, and Holly Lawford-Smith. 2012. "Political Feasibility: A Conceptual Exploration." *Political Studies* 60 (4): 809–825.

Gilbert, Margaret. 2006. *A Theory of Political Obligation: Membership, Commitment, and the Bonds of Society*. Oxford: Oxford University Press.

Goodin, Robert E. 2009. *The Oxford Handbook of Political Science*. Oxford: Oxford University Press.

Harsanyi, John C. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63 (4): 309–321.

James, Aaron. 2005. "Constructing Justice for Existing Practice: Rawls and the Status Quo." *Philosophy & Public Affairs* 33 (3): 281–316.

Joyce, Richard. 2009. "Moral Anti-Realism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2009. http://plato.stanford.edu/archives/sum2009/entries/moral-anti-realism/.

Larmore, Charles. 2013. "What Is Political Philosophy?" *Journal of Moral Philosophy* 10 (3): 276–306.

Leopold, David, and Marc Stears. 2008. *Political Theory: Methods and Approaches*. Oxford: Oxford University Press.

Linz, Juan J., and Alfred Stepan. 1996. *Problems of Democratic Transition and Consolidation: Southern Europe, South America, and Post-Communist Europe*. Baltimore: Johns Hopkins University Press.

List, Christian. 2011. "The Logical Space of Democracy." *Philosophy & Public Affairs* 39 (3): 262–297.

———. 2013. "Social Choice Theory." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2013. http://plato.stanford.edu/archives/win2013/entries/social-choice/.

List, Christian, and Philip Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Lovett, Frank. 2013. "Republicanism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2013. http://plato.stanford.edu/archives/spr2013/entries/republicanism/.

MacCallum, Gerald C. 1967. "Negative and Positive Freedom." *The Philosophical Review* 76 (3): 312–334.

Margolis, Eric, and Stephen Laurence. 2012. "Concepts." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2012. http://plato.stanford.edu/archives/fall2012/entries/concepts/.

McDermott, Daniel. 2008. "Analytic Political Philosophy." In *Political Theory: Methods and Approaches*, edited by David Leopold and Marc Stears, 11–28. Oxford: Oxford University Press.

Miller, David. 1992. "Distributive Justice: What the People Think." *Ethics* 102 (3): 555–593.

———. 2002. "Two Ways to Think About Justice." *Politics, Philosophy & Economics* 1 (1): 5–28.

Mills, Charles W. 2005. "'Ideal Theory' as Ideology." *Hypatia* 20 (3): 165–183.

Mouffe, Chantal. 2005. "The Limits of John Rawls's Pluralism." *Politics, Philosophy & Economics* 4 (2): 221–231.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

O'Neill, Onora. 1996. *Towards Justice and Virtue: A Constructive Account of Practical Reasoning*. Cambridge: Cambridge University Press.

Okasha, Samir. 2002. *Philosophy of Science: A Very Short Introduction*. Oxford: Oxford University Press.

Oppenheim, Felix E. 1981. *Political Concepts: A Reconstruction*. Oxford: Basil Blackwell.

Otsuka, Michael. 2003. *Libertarianism Without Inequality*. Oxford: Oxford University Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Pettit, Philip. 1997. *Republicanism. A Theory of Freedom and Government*. Oxford: Oxford University Press.

———. 2011. "The Instability of Freedom as Noninterference: The Case of Isaiah Berlin." *Ethics* 121 (4): 693–716.

Quine, W. V. 1975. "On Empirically Equivalent Systems of the World." *Erkenntnis* 9 (3): 313–328.

Rawls, John. 1985. "Justice as Fairness: Political Not Metaphysical." *Philosophy and Public Affairs* 14 (3): 223–251.

———. 1996. *Political Liberalism*. New York: Columbia University Press.

———. 1999. *A Theory of Justice*. Oxford: Oxford University Press.

Robeyns, Ingrid. 2008. "Ideal Theory in Theory and Practice." *Social Theory and Practice* 34 (3): 341–362.

Roemer, John E. 1998. *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Ronzoni, Miriam. 2012. "Life Is Not a Camping Trip – on the Desirability of Cohenite Socialism." *Politics, Philosophy & Economics* 11 (2): 171–185.

Sangiovanni, Andrea. 2008. "Justice and the Priority of Politics to Morality." *Journal of Political Philosophy* 16 (2): 137–164.

Sayre-McCord, Geoff. 2011. "Moral Realism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2011. http://plato.stanford.edu/archives/sum2011/entries/moral-realism/.

Searle, John R. 1995. *The Construction of Social Reality*. London: Penguin.

Sen, Amartya K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

———. 1980. "Equality of What?" In *The Tanner Lectures on Human Values, Volume I*, edited by Sterling M. McMurrin, 195–220. Cambridge: Cambridge University Press.

———. 1987. *The Standard of Living*. Edited by Geoffrey Hawthorn. Cambridge: Cambridge University Press.

Simmons, A. John. 2010. "Ideal and Nonideal Theory." *Philosophy & Public Affairs* 38 (1): 5–36.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1 (3): 229–243.

Southwood, Nicholas. 2013. "Democracy as a Modally Demanding Value." *Noûs* Online early: n/a–n/a.

Stemplowska, Zofia. 2008. "What's Ideal About Ideal Theory?" *Social Theory and Practice* 34 (3): 319–340.

Thomson, Judith Jarvis. 1985. "The Trolley Problem." *The Yale Law Journal* 94 (6): 1395–1415.

Tollefsen, Deborah Perron. 2002. "Collective Intentionality and the Social Sciences." *Philosophy of the Social Sciences* 32 (1): 25–50.

Tuomela, Raimo. 2007. *The Philosophy of Sociality: The Shared Point of View*. New York: Oxford University Press.

Valentini, Laura. 2009. "On the Apparent Paradox of Ideal Theory." *Journal of Political Philosophy* 17 (3): 332–355.

———. 2012. "Ideal Vs. Non-ideal Theory: A Conceptual Map." *Philosophy Compass* 7 (9): 654–664.

Van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford: Oxford University Press.

Waldron, Jeremy. 1999. *Law and Disagreement*. Oxford: Oxford University Press.

Walzer, Michael. 1983. *Spheres of Justice: a Defense of Pluralism and Equality*. New York: Basic Books.

———. 1987. *Interpretation and Social Criticism*. Cambridge, MA: Harvard University Press.

Williams, Bernard Arthur Owen. 2005. *In the Beginning Was the Deed: Realism and Moralism in Political Argument*. Princeton, N.J.: Princeton University Press.