Giorgio Bongiovanni · Gerald Postema
Antonino Rotolo · Giovanni Sartor
Chiara Valentini · Douglas Walton *Editors*

Handbook of Legal Reasoning and Argumentation

Springer

Handbook of Legal Reasoning and Argumentation

Giorgio Bongiovanni · Gerald Postema
Antonino Rotolo · Giovanni Sartor
Chiara Valentini · Douglas Walton
Editors

Handbook of Legal Reasoning and Argumentation

Springer

Editors
Giorgio Bongiovanni
Dipartimento di Scienze Giuridiche
and CIRSFID
Università di Bologna
Bologna
Italy

Gerald Postema
Department of Philosophy
University of North Carolina
Chapel Hill, NC
USA

Antonino Rotolo
CIRSFID
Università di Bologna
Bologna
Italy

Giovanni Sartor
Department of Law
European University Institute
Florence
Italy

Chiara Valentini
Department of Law
Universitat Pompeu Fabra
Barcelona
Spain

Douglas Walton
University of Windsor, Centre for Research
in Reasoning, Argumentation
and Rhetoric (CRRAR)
Windsor, ON
Canada

ISBN 978-90-481-9451-3 ISBN 978-90-481-9452-0 (eBook)
https://doi.org/10.1007/978-90-481-9452-0

Library of Congress Control Number: 2018933503

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature B.V.
The registered company address is: Van Godewijckstraat 30, 3311 GX Dordrecht, The Netherlands

Contents

Introduction ix
Douglas Walton

Part I Basic Concepts for Legal Reasoning

Reasons (and Reasons in Philosophy of Law) 3
Giorgio Bongiovanni

Reasons in Moral Philosophy 35
Carla Bagnoli

Legal Reasoning and Argumentation 47
Douglas Walton

Norms in Action: A Logical Perspective 77
Emiliano Lorini

Of Norms 103
Jaap Hage

Values 139
Carla Bagnoli

The Goals of Norms 173
Cristiano Castelfranchi

Authority 191
Kenneth Einar Himma

The Authority of Law 219
Veronica Rodriguez-Blanco

Part II Kinds of Reasoning and the Law

Deductive and Deontic Reasoning 243
Antonino Rotolo and Giovanni Sartor

Inductive, Abductive and Probabilistic Reasoning 275
Burkhard Schafer and Colin Aitken

Defeasibility in Law 315
Giovanni Sartor

Analogical Arguments 365
Bartosz Brożek

Choosing Ends and Choosing Means: Teleological Reasoning in Law 387
Lewis A. Kornhauser

Interactive Decision-Making and Morality 413
Wojciech Załuski

Part III Special Kinds of Legal Reasoning

Evidential Reasoning 447
Marcello Di Bello and Bart Verheij

Interpretive Arguments and the Application of the Law 495
J. J. Moreso and Samuele Chilovi

Statutory Interpretation as Argumentation 519
Douglas Walton, Giovanni Sartor and Fabrizio Macagno

Varieties of Vagueness in the Law 561
Andrei Marmor

Balancing, Proportionality and Constitutional Rights 581
Giorgio Bongiovanni and Chiara Valentini

A Quantitative Approach to Proportionality 613
Giovanni Sartor

Coherence and Systematization in Law 637
Amalia Amaya

Precedent and Legal Analogy 673
Kevin D. Ashley

Economic Logic and Legal Logic 711
Lewis A. Kornhauser

Index of Names 747

Index of Subjects 755

Contributors

Colin Aitken School of Mathematics and Maxwell Institute, The University of Edinburgh, Edinburgh, UK

Amalia Amaya Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México, Mexico City, Mexico

Kevin D. Ashley School of Law and Graduate Program in Intelligent Systems, University of Pittsburgh, Pittsburgh, PA, USA

Carla Bagnoli Dipartimento di Studi Linguistici e Culturali, Università di Modena e Reggio Emilia, Modena, Italy; University of Oslo, Oslo, Norway

Giorgio Bongiovanni Dipartimento di Scienze Giuridiche and CIRSFID, Università di Bologna, Bologna, Italy

Bartosz Brożek Department for the Philosophy of Law and Legal Ethics, Jagiellonian University, Kraków, Poland

Cristiano Castelfranchi Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche (CNR), Rome, Italy

Samuele Chilovi Departament de Filosofia, Universitat de Barcelona, Barcelona, Spain

Marcello Di Bello Lehman College - City University of New York, Bronx, USA

Jaap Hage Faculty of Law, Maastricht University, Maastricht, The Netherlands

Kenneth Einar Himma School of Law, University of Washington, Seattle, WA, USA

Lewis A. Kornhauser School of Law, New York University, New York, NY, USA

Emiliano Lorini IRIT-CNRS Toulouse University, Toulouse, France

Fabrizio Macagno IFILNOVA, Instituto de Filosofia da Nova, Universidade Nova de Lisboa, Lisbon, Portugal

Andrei Marmor Cornell Law School, Cornell University, Ithaca, New York, NY, USA

J. J. Moreso Departament de Dret, Universitat Pompeu Fabra, Barcelona, Spain

Veronica Rodriguez-Blanco School of Law, University of Surrey, Guilford, UK

Antonino Rotolo Dipartimento di Scienze giuridiche, Università di Bologna, Bologna, Italy

Giovanni Sartor Dipartimento di Scienze Giuridiche, Università di Bologna, Bologna, Italy; European University Institute, Florence, Italy

Burkhard Schafer Law School, The University of Edinburgh, Edinburgh, UK

Chiara Valentini Department of Law, Universitat Pompeu Fabra, Barcelona, Spain

Bart Verheij Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands

Douglas Walton University of Windsor, Centre for Research in Reasoning, Argumentation and Rhetoric (CRRAR), Windsor, ON, Canada

Wojciech Załuski Department of Philosophy of Law and Legal Ethics, Jagiellonian University, Krakow, Poland

Introduction

Since ancient times, there has been a presumption by judges, lawyers, and other legal professionals that legal reasoning is based on some kind of rationality that an agent who carries out actions and makes decisions can be presumed to be operating with. For example, legal reasoning does sometimes explicitly appeal to the existence of such a rational agent, called a rational person. But on the other hand, there is much skepticism and controversy about this presumption and more specifically how it applies to legal reasoning. For one thing, there is popular skepticism about whether there is something that can be called legal logic. For another thing, what people traditionally have often seemed to have had in mind is that legal logic fits the model of legal reasoning called mechanical jurisprudence. On this model, deductive logic is used to draw the rational conclusion in a given case at issue, say in a trial, by fitting a strictly universal generalization (All X without allowing any exceptions are Y), to a legal fact X, and drawing a conclusion Y. The problem with this model, although it fits occasionally, it is not applicable to the broad majority of cases in law, where the arguments used to support or attack a conclusion are defeasible (subject to exceptions). So, in the past, we have remained stuck in a dilemma where we are forced to concede that either legal reasoning does not have a logic, or if it does, it is one that is not applicable to the majority of cases being adjudicated on a daily basis.

Recent research on argumentation, especially in the field of artificial intelligence and law, offers a way out of this dilemma, by two means. Argumentation can be defined as a method for identifying, analyzing, and evaluating the pro and con arguments on both sides of a disputed issue where the factual knowledge base needed to resolve a dispute may be incomplete or inconsistent, and fallible arguments are used on both sides to arrive at a provisional conclusion based on a standard of proof appropriate for the case.

Examples of how argumentation tools can be applied to real legal cases are given in this Handbook. One such tool is the use of argumentation schemes, common forms of argument that can be deductive, but for the most part represent forms of

reasoning that are defeasible, as they are subject to criticism or rebuttal by the asking of critical questions. The other is to apply legal reasoning in a dialectical framework which uses burdens and standards of proof, along with other devices, to take the context of use of an argument in a specific setting (e.g., in a trial in a particularly legal system) into account. Moving forward with this task means that linguistic interpretation of legal terms needs to be treated as inherently pragmatic in nature. Such an approach must not only take into account the semantic meaning of words and expressions, but also their pragmatic aspects considering how they are used in a communicative context. This latter approach requires not only considering the rationality of both a single agent with individual goals, but also the rational decisions and actions of several agents who reason together to deliberate to carry out their collective goals.

This Handbook shows, from a number of different angles and perspectives, and using a number of different tools, many of which may be new to readers, how this new approach to legal reasoning can be applied to many different important aspects of legal reasoning, throwing light on number of key problems and providing new avenues for solving them. By this means, the reader is allowed to look at legal reasoning in a fresh way, and thereby move forward to overcome the traditional dilemma about whether there is a legal logic or not.

One of the problems confronting the traditional approaches to legal reasoning is the uncertainty among legal scholars at this point in time, about the relationship between argument, at least argument in the sense of the term representing rationality, and reasoning. As one might expect, at this time there are also differences among the theoreticians on how to define the notion of an argument in precise enough terms to make the concept useful for the study of computational models of legal argument. Moreover, as noted above, accepting standard models of reasoning that have been dominant in the past, such as those of classical deductive logic, represents legal reasoning as mechanical jurisprudence based on absolutely universal rules of law not subject to exceptions. Hence, the problem of how to define the notion of an argument in a way that enables the drawing of distinction between reasoning and argument is one that pervades attempts to model legal reasoning by some more flexible notion of argument that can do justice to defeasible legal inferences of the most typical kind. Many of the chapters in the Handbook confront this problem.

Although some theorists now prefer the language of arguments, more traditional theorists would like to have a point of entry into this family of concepts by first of all defining what seems be the less problematic notion of a reason. Giorgio Bongiovanni, in Part I, Chapter "Reasons (and Reasons in Philosophy of Law)," provides a theory to answer this question and to classify different kinds of reasons. In this chapter, he distinguishes between normative, motivating, and explanatory reasons, and provides a theoretical framework for drawing a distinction between reasons for belief and reasons for actions. This chapter helps the reader move forward to the other chapters based on the assumption that some sense can be made of the distinction between the fundamental idea of presenting reasons to reasonably accept or reject a disputed claim, and the idea of presenting pro or con arguments for this

same purpose. Overcoming this linguistic and underlying theoretical problem paves the way for the explorations of legal reasoning in the rest of the book.

Reasons in law are of course closely related to reasons in moral philosophy. Carla Bagnoli, in Chapter "Reasons in Moral Philosophy" of part I, clarifies the functions of moral reasons by drawing a distinction between explanatory reasons that make an attitude or action intelligible, and normative reasons that guided an agent's activity by offering considerations in support of or against actions. Even though this distinction by Bagnoli is not taken to be mutually exclusive, it provides a pragmatic basis for understanding how reasons have rational bite in different contexts where reasoning is typically used. It also offers a basis for distinguishing between subjective and objective reasons, which is useful for understanding how one agent can have authority over another, based on the assumption that normative reasons can be founded on arguments from authority. These distinctions are relevant to legal reasoning because they can lead to a better understanding of norms of basic rationality that can help a citizen to deal with conflicts of moral reasons, a kind of problem widely confronted in legal argumentation. This problem raises the question of how such common legal conflicts can be resolved in a legal setting by logical arguments. For example, they raise the question of how evidence-based reasons can help legal adjudicators decide outcomes in cases where there are moral reasons on both sides of a conflict.

Chapter "Legal Reasoning and Argumentation" of part I shows how the arguments on both sides of a case can each be based on evidence, such as witness testimony and so forth, that can support rational arguments and that in themselves can represent rational arguments. This treatment of such arguments by Douglas Walton in the chapter lends support to Wigmore's view that there is some kind of science of proof apart from deductive logic that underlies legal reasoning. The examples treated in the chapter show how such arguments are evidence-based. This basis in evidence gives us a structure for analyzing and evaluating how legal argumentation works in general (and, more relevantly for this Handbook, should work) as applied to particular cases. The analyses of the examples of legal reasoning in this chapter show us how to apply typical defeasible argumentation schemes, such as argument from witness testimony, argument from expert opinion, abductive reasoning, and so forth, to arguments put forward on either side of a contested case. It also shows us how to analyze and evaluate sequences of argumentation by chaining together such individual arguments based on schemes in a context such as that of a trial. Additionally, Chapter "Legal Reasoning and Argumentation" shows how the argumentation in such a sequence has three stages, an opening stage, an argumentation stage, and a closing stage. The middle part, that of the individual arguments making up the chain of pro and con argumentation, represents the reasoning used in a case, whereas the other two parts represent argumentation in the fuller the dialectical (procedural) sense of the word. They provide essential parts of the pragmatic aspect of the argumentation, taking us from the burden of persuasion at the opening stage to the decision made during the closing stage.

The concept of an autonomous rational agent carrying out an intelligent goal-directed action is fundamental to computing, especially in multiagent systems and robotics, and to understanding legal reasoning and argumentation. Yet little has been done to apply work on action theory to legal reasoning and argumentation in trials and other legal settings. In Chapter "Norms in Action: A Logical Perspective," Emiliano Lorini provides a clearly written survey and introduction covering some most important results in this field that can be applied to legal reasoning. Drawing on rich historical sources, and the formally developed logical systems of norm and action that can be found in the writings on action by logicians and philosophers, Lorini explains the state of the art on the most promising development in this area, the so-called *STIT*, or the logic of "seeing to it that." This logical model represents the basic idea of a rational agent bringing it about that a particular proposition is true or false by means of carrying out an action. The underlying idea is that this concept can be modeled as an agency operator in a modal logic system using a Kripke-style semantics. The formal semantics of *STIT*, which is both elegant and intuitively understandable, can be applied to almost any example of legal argumentation of the kind found typically in the courts in any jurisdiction. It offers a logical structure for framing reasoning about choices, actions, and time that is easily applicable to the evidential reasoning in legal cases.

The logical formalization of *STIT* has become an intricately built framework in recent years, and Lorini provides the service of presenting an outline of the main results and applications of the system, showing how they can be applied to the formalization of such key legal notions as responsibility and influence. It is shown, for example, how *STIT* can be applied to the type of responsibility that consists in one agent inducing another agent to violate a certain norm so that the influencer becomes indirectly responsible for the norm violation, and is subject to a sanction. This connection leads to Chapter "Of Norms," which is on the role of norms in legal reasoning.

Norms are, to put it briefly, social and/or legal requirements that separate actions into three categories, those that are required (obligatory), those that are permitted, and those that are forbidden (prohibited). Chapter "Of Norms," by Jaap Hage explains that there are different theories of norms, and different usages of the word "norm" in English, and corresponding terms in other languages. A norm is sometimes described as a prescriptive guide to acting as a command that empowers, proscribes, or allows actions. Obviously, norms are very important in law and ethics. Recently, study of norms has become important in the development of multiagent systems in artificial intelligence, and new formal argumentation systems incorporating norms have thrown light on how reasoning is based on norms in law, ethics, and other applications.

In particular, Hage distinguishes between two kinds of norms, one that tells us what to do, and one that informs us what ideally should be a case. Hage concentrates on norms that have the function of guiding human behavior, clearly a mainstream concern in legal reasoning. He shows us how norms are closely related to reasons for action and clarifies the distinction between norms and facts by distinguishing among various kinds of facts, the way this term is conventionally

used. Hage explains how norms are related to what are called possible worlds in the standard semantics for modal logic. He explains how norms are related to duties and obligations. He clarifies the notion of a norm by defining it as a rule that leads to deontic consequences. This chapter is fundamentally important for understanding legal reasoning because it brings out not only how norms are fundamental to legal reasoning, but also how our reasoning with norms can be modeled by the form of modal logic called deontic logic. In Chapter "Deductive and Deontic Reasoning" of part II, Antonino Rotolo and Giovanni Sartor present an introduction to deontic reasoning, the logic of norms.

What are values, and how can we identify them in particular cases? These are the questions addressed Carla Bagnoli, in Chapter "Values," shows how theories of value answer questions such as how we can judge the adequacy of a theory of value, and how values can have a normative capacity to guide action. She shows how theories of values offer answers to these questions, and can throw light on some disagreements in legal cases that depend on arguments about the incommensurability of values.

Values have become especially noticeable as a current topic in artificial intelligence and law now that it has been shown that practical goal-directed reasoning is not always purely instrumental, but is commonly based on values as well, especially in a legal setting. However, philosophical questions remain about the nature of values, and how they play an especially important role in legal reasoning.

Chapter "The Goals of Norms," authored by Cristiano Castelfranchi, is about the relationship between norms and goals. One immediate connection that he identifies at the outset and brings out in this chapter is that norms are artifacts for social coordination through a rational agent's manipulation of its own or others' goals (whether the agent is a machine or a human). This interconnection is shown by Castelfranchi to go both ways. Norms are used in the legal reasoning of goal-directed agents whose actions depend on their free decisions, but norms also have goals and that they are built by agents and used for something. They are societal tools. Norms depend on goal-directed practical reasoning because they have the function of trying to actuate the intended effects corresponding to goals. Goal-directed practical reasoning is often called teleological reasoning, referring to an agent's purpose in carrying out an action. One does not have to go very far into this network of concepts to appreciate how fundamental they are to understanding legal reasoning.

Castelfranchi lists six main structural relations between norms and goals. First, norms are designed to influence autonomous goal-directed actions of a rational agent. Second, norms presuppose the postulation of goals in the mind of an agent carrying out an action. Third, norms are aimed at governing our conduct and often give us new reasons for or against a goal. Fourth, norms are internalized and adopted for a goal that an agent has. Fifth, norms have goals, because they are aimed at bringing about certain outcomes. Sixth, norms are based on collective expectations about the goals of other agents.

An especially helpful part of this chapter on the use of the term "goal" in legal reasoning is that it explains how in modern science there are two different

theoretical approaches to the concept of a goal. One is provided by evolutionary approaches, while the other is provided by the control theory of cybernetics. Bringing these two approaches together and explaining both of them is very helpful for seeing where the study of goal-based practical reasoning in artificial intelligence and law is going. Goals, motives, and intentions are fundamental notions in legal reasoning, especially in criminal law. By showing how these concepts are related, and in turn related to norms, this chapter throws considerable light on fundamental concepts of legal reasoning. In legal reasoning, goals are typically based on values, producing the kind of reasoning called value-based reasoning, as opposed to purely instrumental reasoning.

Authority has long been recognized as an important concept for law, but now with the recent literature in artificial intelligence on evidential reasoning based on authority, such as expert opinion evidence in trials, the concept of authority has become more important than ever for understanding how legal reasoning works. In Chapter "Authority," Kenneth Einar Himma draws a distinction between two kinds of authority. Epistemic authority is the source of reasons to believe that a proposition is true or false, acceptable or not, based on the evidence. Practical authority is the source of reasons for action. This chapter is concerned with the notion of practical authority. The chapter identifies properties that make something of practical authority, explains the kinds of reasons that bind subjects of this kind of authority, and examines what conditions standards of practical authority must satisfy to be morally legitimate.

In Chapter "The Authority of Law," Veronica Rodriguez-Blanco poses the philosophical question of how a person as a rational agent can be in control of her own destiny, given that law requires us to carry out innumerable actions which we freely and intentionally perform all the time. To approach this problem, she focuses on the agent and works upward from the practical reasoning of an individual agent to the framework of authority. Instead of trying to explain human actions as being exclusively empirical phenomena, she perceives the need to understand human action in a more fundamental form, seeing it as it operates in a framework of human institutions such as law. This ties in with the need for paying further attention to action theory by applying logical models such as STIT to study the notion of a rational agent carrying out an intelligent goal-directed action, concept one that is fundamental to both computing and legal reasoning.

Drawing on the philosophical literature on intention, she points out that intentional action involves knowledge that is not of an observational kind, even though it might be expedited and supported by observations. On this approach, understanding an action according to reasons or intentions should begin by way of asking a why-question. On this way of viewing an explanation of a person's actions, we grasp it from the person's own description of his action given as an answer to a why-question. This dialectical multiagent approach to the evidential basis for reasoning to intentions is compatible with recent work on legal argumentation.

In Chapter "Deductive and Deontic Reasoning" of part II, Antonino Rotolo and Giovanni Sartor present an introduction to deductive and deontic reasoning, as these formalisms apply to legal reasoning. Since there are already many formal

systems of deontic logic, this chapter starts with deductive logic. The formal systems of deontic logic took were developed a framework of deductive logic (classical logic), so to understand much about formal systems of deontic logic, you have to start with deductive logic. But from there, the chapter offers an account of the basics of deontic logic that uses simple examples that are easily transferable to a legal context, so that the reader can easily appreciate and understand how it works in law.

The chapter explains how statements about obligations and permissions work as modal operators in a classical system of deductive modal logic and how such systems apply to conflicts of obligations and permissions of the kinds familiar in law. The chapter outlines the basics of the Kripke semantics as it applies to deontic logic, and from there outlines some axioms and theorems in the system as they apply to common logical inferences of the deontic kind using ordinary legal examples. The logical system is shown to be sound and complete. An interesting feature is that the chapter explains why more advanced normative notions, such as the notion of a right, cannot be exclusively built on the basis of obligations and permissions, because rights can only be analyzed by making reference to the interests of an agent.

Now that Chapter "Deductive and Deontic Reasoning" has covered how deductive logic applies to legal reasoning, Chapter "Inductive, Abductive and Probabilistic Reasoning" proceeds to investigate the role of probabilistic reasoning in law. Probability is always a difficult and contestable subject on the issue of how it applies to legal reasoning, especially when those of us without specialized knowledge of probability theory and the Bayesian axioms try to apply them to real arguments. By using common legal examples to illustrate how it works, and how examples of its application have been subject to interpretation and controversy, Chapter "Inductive, Abductive and Probabilistic Reasoning" is especially valuable for those of us who are uncertain about just how far Bayesian probability can go in analyzing and evaluating the kinds of evidential reasoning commonly found in trials.

In Chapter "Inductive, Abductive and Probabilistic Reasoning," Burkhard Schafer and Colin Aitken survey the history of the relationship between probabilistic reasoning and jurisprudence, showing how the emergence of the subjectivist view of probability has come to be of pivotal importance. On this view, probability represents the subjective degree of belief of a proposition. This view has turned out to be particularly important for legal reasoning because the inferences drawn by jurors are based on background knowledge and common sense assumptions that are difficult to reduce to objective statistical propositions.

As things have turned out, the most widely explored route to try to find a satisfactory application of probability to legal reasoning in a broad majority of cases has been the subjective Bayesian approach based on Bayes' theorem. There have been many differences of opinion on how to apply Bayes' theorem to areas of legal reasoning, such as evidence based on witness testimony, or the arguments from precedent and counter-arguments attacking these arguments. By using a number of relatively clear and simple ordinary examples throughout this chapter, Schafer and

Aitken explain how probabilistic reasoning and abductive reasoning, the latter usually associated with inference to the best explanation, relate to ongoing jurisprudential debates and match forms of argument commonly used in legal reasoning.

In Chapter "Defeasibility in Law," Giovanni Sartor surveys the leading formal and computational theories of defeasibility that have been prominent in artificial intelligence and law, but at the same time shows how the idea of defeasible reasoning can be traced back to Aquinas and Leibniz, and even to Cicero. Sartor explains how the process of defeasible reasoning reflects the natural way in which legal reasoning proceeds, given that law is applied to particular situations, typically in cases where conflicting legal rules may apply, so that adjudication must work with conflicts between the rules as they apply in a case. This chapter approaches defeasible reasoning from an argumentation point of view, where the evidence is evaluated by a process of critical questioning along with weighing pro and con arguments that are relevant within a framework where there is a burden of persuasion is to decide the outcome.

Sartor visually represents the argumentation in a number of interesting examples of legal reasoning by using argument diagrams. The examples, once analyzed, show some interesting features of legal argumentation. One is that when new arguments are introduced into what is called an argument framework representing the set of arguments constructible from a given set of premises, the status of a given argument can change relative to that framework. For example, an argument that was previously justified in the framework can now be overruled. Structuring argumentation in this way, an argumentation framework can be used not only to evaluate a complex sequence of argumentation taking the form of an argument graph or argument diagram. It can also be used to extrapolate the argumentation forward to support or attack an ultimate claim to be proved. By this means, Chapter "Defeasibility in Law" provides an overview of how argumentation systems provide dialectical frameworks representing the pragmatics of legal reasoning as well as its semantics. This capability is fundamental to understanding how legal reasoning works in a case-based setting.

In Chapter "Analogical Arguments" (*Analogical Arguments*), Bartosz Brożek begins by surveying the topic of argument from analogy from Greek philosophy to the recent tradition in philosophy of science and psychology that portrays analogy as a kind of cognition. To illustrate uses of argument from analogy in science, everyday conversational reasoning and legal reasoning, Brożek analyzes a series of examples that shows how analogical reasoning works, taking an argumentation approach in which problems give rise to certain kinds of questions, notably in some cases open-ended questions. Taking a formal approach, he shows how, once the problem situation has been identified, logical argumentation proceeds by retrieving a set of previously decided cases that are similar to the problem situation in certain respects.

From that point, the chapter applies the formal analysis to some well-known legal cases such as the case of Adams v. New Jersey Steamboat Company. Of special interest to the readers of this Handbook is the analysis of how relevant

similarity works as a key component in legal analogical arguments. This is shown by using the same extended legal examples and fitting them to the theoretical framework for argument from analogy. It is shown that there are two widely accepted general methods of evaluating legal arguments from analogy, one called the theory-based approach and the other called the factor-based approach. These two approaches are combined into a general structure for analogical arguments that can be applied to cases legal argumentation, L as illustrated by the examples analyzed in the paper. One thing that comes out clearly in the chapter is the need to take the dialectical dimension of analogical arguments into account in order to adequately model their uses in legal reasoning.

In Chapter "Choosing Ends and Choosing Means: Teleological Reasoning in Law," Lewis A. Kornhauser provides a theory to explain the process of teleological reasoning by articulating its nature in relation to rational choice theory. Teleological reasoning is basically goal-directed reasoning by a rational agent who could, for the purposes of this Handbook, be either a machine or a human, and can comprise a group of agents forming a team for the purpose of deliberating on what to do. Teleological reasoning is identified in argumentation studies by the argumentation scheme for practical reasoning, stating that if an agent has a goal, and knows of the means to carry out that goal, then other things being equal, the agent should go ahead and carry out the action that is the means. Teleological reasoning, in this sense, is a defeasible kind of argumentation subject to default if critical questions are asked when new information comes in, such as the question of whether alternative means are available, or the question of whether carrying out the action would have negative consequences for the agent. According to Kornhauser's account, a rational agent must pay attention to what aspects of consequences are involved as well.

One important feature of teleological goal-directed legal reasoning according to Kornhauser is that legal goals do not need to always be moral ones, meaning that they have to be based on values as well as goals. An important observation made is that legislatures often have to operate instrumentally when they decide which legislation to enact and promote through statutes. An example he cites is the enactment of the Clean Air Act by the US Congress on the basis of the reasoning that the passing of the act will have the consequence of reducing pollution. Cases like this show a process of stepwise reasoning from the actions of one agency to those of another. For example, one institution sketches a goal, a second institution elaborates the goal, and a third agent has the task of implementing the goal, say by framing and implementing a law.

It is becoming more and more obvious in artificial intelligence and law how teleological reasoning is both widespread in legal institutions and how it is fundamentally important generally for understanding how legal reasoning works in practice. But there is already such an extensive literature on consequentialist theories of value and ethics and on rational choice theory as a framework of goal-directed decision-making that it is intimidating to most readers without a background in these areas to get any clearer idea of what teleological reasoning is and how it applies to law and legal decision-making. This chapter uses simply

explained examples that guide the reader through this bramble bush of interrelated writings and theories with the clarity that enables him or her to see what elements of them can be helpful to clarify legal teleological reasoning.

There is a special type of decision-making called interactive decision-making which takes place in multiagent settings where what each agent does is dependent on not only with the other agents do but also on their expectations of what the other agents in the decision-making group can be expected to do. Wojciech Załuski, in Chapter "Interactive Decision-Making and Morality," explains how interactive decision-making called strategic decision-making in game theory can contribute not only to moral philosophy but also to our understanding of legal reasoning as a goal-directed form of argumentation.

He begins by outlining the assumptions put to work in classical game theory, showing how there can be stronger and weaker assumptions about the knowledge that the players have and the degrees of rationality that they and the other players can be assumed to have. This strategic approach assumes that the players can make mistakes, and that each player can take advantage of the mistakes made by the other players. This approach has advantages for a good fit to analyzing legal reasoning in an adversarial system, bringing out important aspects of the argumentation that traditional theories of legal reasoning often tended to overlook or minimize. Załuski presents some basic examples of classical game theory that throw light on these aspects. It is also shown in this chapter how game theory can work as a tool for criticizing moral assumptions and theories of the kinds often applied in legal argumentation. Particularly important is the distinction between instrumental teleological rationality and value-based teleological rationality, notions that permeate legal reasoning and deliberation.

Chapter "Evidential Reasoning" of Part III is a general survey of the main problems of evidential reasoning that have been studied in artificial intelligence and law, and explains the leading theories that have been proposed as a way of solving these problems or moving ahead to provide a more unified account of legal reasoning by connecting the theories or even merging them. An attractive feature of this chapter is that even someone with a limited background in artificial intelligence or logic, or related technical subjects, can apply it to understanding how the nuts and bolts of legal and logical reasoning are put together. Common examples of legal reasoning are used, mainly from criminal law, and argument diagrams are presented so that the reader can visualize the basic structure of the reasoning in each case easily and clearly. In each case, an explanation enables the reader to understand how tools from artificial intelligence can be applied.

Di Bello and Verheij cover such basic kinds of evidence as witness testimony evidence, and scientific expert testimony evidence, such as DNA evidence of the kind that has now become so common in criminal cases. They go on to how we should understand conflicts between pieces of evidence, how we should evaluate strength of the evidence, how we should interpret the available evidence, how we should decide about the facts given the evidence, and it should be decided that an investigation has been exhaustive enough so that the closing stage of a criminal proceedings is reached.

An extremely useful part of the chapter is the outline of the three normative frameworks that have been put forward as systematic and well-regulated methods for examining, analyzing, and weighing the evidence in a case. These are the argumentation framework, the probability-based framework, especially Bayesian methods, and the scenario framework that sees the determination of the outcome and a legal case, for example in a criminal trial, as an argumentation-based rational decision between competing stories. The various leading problems in applying these normative frameworks to everyday legal argumentation are explained and discussed, and suggestions are made on how to solve them.

Chapter "Interpretive Arguments and the Application of the Law," by J. J. Moreso and Samuele Chilovi, surveys the literature on current theories of how to interpret the law, addresses criticisms of them, and present their own theory. The first theory considered is the communicative content theory of law, which holds that legal interpretation is modeled on utterance interpretation. This view holds that facts about the nature of language, taken together with facts about the nature of law, are all that is needed to drive a legal interpretation as a conclusion and tell whether it is correct. However, Moreso and Chilovi hold that there is a conflict between this theory and the doctrine of the rule of recognition which suggests that the communicative content theory is problematic.

The next theory considered is the communication theory of law, which holds that legal content is determined in the same way that propositions and other elements of linguistic texts are interpreted in ordinary language. This theory uses what is called a principle of epistemic asymmetry according to which the producer has a message he wants to get across is a particular form of words, and the consumer operates on the assumption that the producer meant something. If the consumer interprets the producer correctly, then the consumer is taken to have succeeded in identifying what the producer meant.

The difficulty with the communication theory of law, according to Moreso and Chilovi, is that it requires the consumer to select among the various intentions that the speakers might have and select the one that is relevant for the legal application. The problem with this approach, they contend, is that it remains unclear what the object of interpretation precisely is. The difficulties are that there is the Gricean problem of meaning something without saying it, using implicature. Another problem is that there can be a communication failure where the rational here takes the speaker to have meant something different from what she actually said. This is shown in detail by an examination of the Gricean maxims as they might be applied to problematic cases of legal interpretation.

After a close examination of these theories, resting on a series of examples, Moreso and Chilovi, put forward their own theory as an alternative that, they claim, can reply to all the objections they encountered in treating the existing theories. According to the theory of Moreso and Chilovi, the existing legal theory already has an assemblage of types of arguments, such as argument from analogy and argument *a contrario*, that can be applied to norms to facts to generate an interpretation (such as one of a statute) using only deductive reasoning. From this premise, they conclude that no form of logic other than classical deductive logic is

needed to provide a logical structure to underpin their theory of interpretation, except for an extension of deductive logic to deontic logic.

A valuable feature of this chapter is that it reveals significant weaknesses in the leading traditional theories of legal interpretation, and thereby provides an interesting survey of the theories themselves in the difficulties inherent in them that would enable the study of statutory interpretation to move ahead.

In Chapter "Statutory Interpretation as Argumentation," Douglas Walton, Giovanni Sartor and Fabrizio Macagno show how the traditional canons of interpretation can be represented as argumentation schemes that are defeasible forms of argument pro or con the interpretation of a given statutory or legal text. The formalization of the schemes given in the chapter makes possible the modeling of legal interpretation using formal argumentation systems from artificial intelligence. After introducing some of these formal systems and applying them to two cases, the chapter develops a logical model for reasoning with interpretive canons from a text using defeasible rules to draw an interpretive conclusion.

The chapter begins with a list of eleven interpretive arguments, including argument from ordinary meaning, argument from technical meaning, argument from contextual harmonization, argument from precedent, argument from analogy, argument from a legal concept, argument from general principles, argument from history, argument from purpose, argument from substantive reasons, and argument from intention. The names themselves roughly indicate the nature of each type of argument. This list of eleven types of argument which can be used to support or attack a legal interpretation is compared to an overlapping list of fourteen types of arguments previously identified in the literature. The chapter shows how these interpretive arguments can be classified into subtypes and how each of the schemes representing them need to be formulated so that in this form can be used to derive interpretations in several key examples. These interpretative schemes provide ways of dealing with vagueness and ambiguity in law.

In Chapter "Varieties of Vagueness in the Law," Andrei Marmor distinguishes between different kinds of vagueness in law and explains some of the ways in which legal decision-makers reason with the language. Slippery slope arguments of the kinds one finds in law occasionally, sometimes turn out to be very controversial because they depend on the vagueness of a key legal term. Vagueness, or open texture as it has been called in legal contexts, is inevitable both in ordinary language and in legal communication and reasoning. Legal reasoning itself takes place in natural language, and so as Marmor shows, law cannot entirely avoid linguistic vagueness, even though it has ways of dealing with it. Terms such as "reasonable care," "due process," and so forth, can be made more precise for legal purposes by precedents and criteria set by law, but the inherent vagueness in them is unavoidable because all natural language terms are open-textured. There are always going to be borderline cases. Vagueness is something that case-based legal reasoning can deal with, and it has to contend with on an ongoing basis.

Marmor draws a distinction between semantic vagueness, which concerns the relations between the meanings of words and the objects they apply to, and conversational vagueness, which has to do with borderline cases and relevance. Both

kinds of phenomena occur in legal reasoning. As Marmor shows, the normal procedure in law when regulating with vague standards is to put the decision for sanctions for violation to the courts so they can decide whether the standard was violated or not in a given case. By this means, the precedent meaning set by the courts makes the standard less vague in a certain respect.

Marmor shows that this procedure of using legal reasoning to make a standard more precise is context sensitive, and hence is a matter of pragmatics in linguistics (the study of meaning that takes contextual factors into account when drawing implications about what a word or phrase may be taken to mean in a specific instance of its usage). A pragmatic approach is necessary for statutory interpretation, as shown by Walton, Sartor, and Macagno in Chapter "Statutory Interpretation as Argumentation."

In Chapter "Balancing, Proportionality and Constitutional Rights," Giorgio Bongiovanni and Chiara Valentini explain how and why proportionality review is a widespread decision-making model that lies at the core of the debates on rights in education, where it has raised questions about the nature and distinctive features of legal reasoning. In this chapter, they examine the relation of different forms of proportionality to the balancing of rights. This chapter explains how conflicts of interests which lie at the foundation of rights illustrate the need for legal reasoning of a kind that has the capability to distinguish between principles and rules.

The chapter reviews several leading theories that propose models of constitutional rights, revealing that they show the need to apply the canon of proportionality, requiring an approach in which laws operate on different hierarchical levels. Such an approach is shown to require an account of value-based legal reasoning whereby value can be based on the interests of an agent. On this view, if the agent has an autonomy interest in an activity, that activity must be protected by a right. The most influential model of proportionality–balancing holds that constitutional rights need to be treated as defeasible principles that may conflict with other rights or interests, where such conflicts need to be adjudicated by a process of optimization. Taking this approach, it is shown how justification of proportionality review is the legal instrument that has been and needs to be adopted by the courts. It is shown how proportionality takes two fundamental forms, optimizing proportionality and state-limiting proportionality. Several other alternative approaches to proportionality–balancing are considered as well.

In Chapter "A Quantitative Approach to Proportionality," Giovanni Sartor addresses the extent to which the operations involved in balancing and proportionality assessments may include quantitative reasoning, and be subject to arithmetic constraints. Relying on some work on cognitive and evolutionary psychology he argues that processing non-symbolic approximate continuous magnitudes is a fundamental cognitive capacity, which seems to be deployed also when we are reasoning with values, as scalable goals are being pursued. A model is proposed for determining the impact of a choice on different values, assessing the utilities so produced and merging these utilities into an overall evaluation, which may be used in comparisons. The usual standards deployed in proportionality assessments, such as suitability, necessity, and proportionality in a strict sense, are specified relatively

to this model. Finally, it is discussed how proportionality assessments can lead to the formulation of rules, and how quantitative proportionality assessments may be constrained by the requirement of consistency with precedents.

Coherence has recently been revived as an alternative to foundationalist theories of justification and as an alternative to the Bayesian model of reasoning, both in psychology and philosophy. Coherence has also been appealed to legal theory as a standard of rational justification. The three main kinds of theories of normative coherence are reviewed and evaluated: principle-based theories, case-based theories, and constraint-satisfaction theories. In Chapter "Coherence and Systematization in Law," Amalia Amaya focuses on normative coherence as dependent on legal coherentism. She addresses the problem of coherence bias, formulating this problem in detail, and argue that a modified version of coherence bias can address this problem. The chapter comparatively evaluates the value and limits of coherentist reasoning in law.

In Chapter "Precedent and Legal Analogy" of part III, Kevin D. Ashley shows how argumentation schemes representing argument from analogy and arguments from precedent can be applied to structure arguments employed in court opinions whether or not to apply precedent or a legal analogy in specific cases. This exercise is valuable for helping law students, legal professionals, and theorists of legal reasoning to both support arguments and to attack them in a carefully reasoned manner. Chapter "Precedent and Legal Analogy" builds on recent work in argumentation. Artificial intelligence has provided a repository of argumentation schemes that can provide a prima facie reason tentatively accepting the conclusion of an argument based on the acceptability of its premises. Such schemes can be used to link arguments together, so that a connected network of such arguments can be evaluated by weighing the pro arguments against con arguments using formal argumentation systems from artificial intelligence.

Chapter "Economic Logic and Legal Logic," written by Lewis A. Kornhauser, compares legal reasoning to the kind of reasoning used in economics and uses this comparison to argue that the latter deepens our understanding of the former. He shows how economic models abstract from the details of complex social phenomena. He suggest that for this reason, it is not surprising that economic conclusions are generally arrived at using mathematical models that rely on deductive reasoning and statistical probability. They have also tended to require that each agent in the decision-making situation as a complete ranking of the outcomes. Therefore, although economic reasoning, like legal reasoning, is based on goal-directed means-end reasoning, it typically concentrates on the means and takes the agent's goals as given. For these reasons, economists have tended to shy away from value-based practical reasoning and do not take used to take the rational agent's values into account.

However, behavioral economists now recognize that agents systematically deviate from these strict rationality assumptions, and this departure presents prospects for the application of argumentation to the study of our reasoning takes place when a (somewhat) rational agent carries out an action such as choosing to buy or sell something. Argumentation would take the approach that this procedure works

by the agent deliberating on both sides of the issue using pro–con argumentation, where the sequence of argumentation on both sides is based on argumentation schemes. Argumentation schemes can be deductive or probabilistic in the statistical sense in some instances, but as shown in this Handbook, studies in argumentation have recently shown that in cases of this kind, deciding whether or not to buy something for example, agents use defeasible argumentation schemes, such as the scheme for goal-directed practical reasoning and the scheme for argument from negative consequences. For these reasons, judging from what Kornhauser has shown in this final chapter of Part III, there is a brave new world out there ready to explore how argumentation applies to economic reasoning.

Douglas Walton

Part I
Basic Concepts for Legal Reasoning

Reasons (and Reasons in Philosophy of Law)

Giorgio Bongiovanni

1 Premise

Notwithstanding the fact that reasons have been at the centre of the reflection on normativity and action for at least forty years,[1] there is not complete agreement about the concept and the features of reasons. What is a reason, what kinds of reasons there are, what their different qualifications are, and so on are the subject of a wide discussion.[2] As noted by J. Searle (referring to P. Foot)[3] what is a reason for action seems "to be frightfully difficult," even though "we deal with reasons for action every day" (Searle 2001, 97).[4] This difficulty has different sources, ranging from the "heterogeneity in the use of the term 'reason'" to the different meanings we can attribute to it. As Alvarez (2010, 8) notes, "it is common," in dealing with these questions, "to introduce the discussion by drawing a distinction between different *senses* of the term 'reason,' or between different *kinds* of reason."[5] But distinguishing between reasons can be done in different ways, with different points of reference,

[1] Broome (2004, 28) notes that "within the philosophy of normativity, the 1970s was the age of the discovery of reasons."

[2] For Alvarez (2010, 1), the analysis of reasons raises the following questions, among many others: "What are reasons? Are there different kinds of reasons? Are reasons beliefs and desires? If not, how are they related to beliefs and desires? And what role do they play in motivating and explaining actions?"

[3] Searle (2001, 97) relates that Philippa Foot once wrote, "I am sure that I do not understand the idea of a reason for acting, and I wonder whether anyone else does either."

[4] Dancy (2000, 1), underlines that "There are not so many things that we do for no reason at all. Intentional, deliberate, purposeful action is always done for a reason, even if some actions, such as recrossing one's legs, are not—or not always, anyway."

[5] Alvarez (2010, 7) notes that "this territory is […] quite complex. And with it comes the temptation to suppose that the machinery required to find one's way around it must be correspondingly complex."

G. Bongiovanni (✉)
Dipartimento di Scienze Giuridiche and CIRSFID, Università di Bologna, Bologna, Italy
e-mail: giorgio.bongiovanni@unibo.it

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_1

and it is not easy to find a common denominator.[6] A possible and general initial point is that of differentiating reasons according to their roles in action and reasoning. This analysis, which starts from the awareness that the same reason may play different roles in different contexts,[7] seems to be prodromic to the classification of reasons, it does not commit one directly to an ontological point of view (although it is functional to a unified view of reasons), and it enables us to "explore reasons broadly" (and after focusing on the different kinds of reasons). In this way, this analysis appears useful, as it makes it possible to provide a first classification of reasons and specify the different questions that an analysis of reasons poses, and, at the same time, it could make it possible to order the questions on a different level of generality.

The following sections will analyse (a) the definition of the different classes of reasons (normative, motivating, explanatory) in relation to their role; (b) the problem of the ontology of reasons (facts or mental states); (c) the kinds of reasons and in particular the distinction between reasons for belief (epistemic) and reasons for action (practical); (d) the modality and the strength of reasons; and (e) reasons and the law.

2 The Different Classes of Reasons: Normative, Motivating, Explanatory

Reasons play different roles in our life: even at a glance, we can see how they can be used to guide, motivate, justify, understand and evaluate our behaviours and believes. From this perspective, it is possible to note an assortment of roles that reasons can play, namely, to "motivate and guide us in our actions (and omissions)"; to "be grounds for beliefs, desires, emotions, etc. […] to evaluate, and sometimes to justify, our actions, beliefs, desires, and emotions"; and to be "used in explanations" (Alvarez 2010, 7).[8]

A general distinction often introduced in contemporary literature to summarize these different roles is that between *normative* and *motivating*[9] reasons: the former are the kind that, as is well known, favours or "counts in favour of" a specific behaviour or acting in a specific context,[10] while motivating reasons are the kind "for which

[6] Alvarez (2010, 10) stresses, in this sense, that there are "different ways of partitioning."

[7] Alvarez (2016a) argues that "in itself and out of context, a reason is not a reason of any particular kind, say normative or motivating. It is only in a particular context, where the reason plays a specific role and can be cited to answer a particular question that it can be qualified as being of this or that kind."

[8] In the same sense, Raz (1999a, 15–16) notes that "as well as reasons for actions there are reasons for beliefs, for desires and emotions, for attitudes, for norms and institutions, and many others […]. Reasons are referred to in explaining, in evaluating, and in guiding people's behaviour."

[9] Sometimes, the distinction is placed between *normative* and *explanatory* reasons. See Bagnoli, chapter 2 , part I, this volume, on "Reasons in Moral Philosophy," para 2.

[10] Of course, this idea of reasons as "favouring" comes from Raz (1999a) and Scanlon (1998). Broome (2004, 41), who sees as "a commonplace" the idea that "the reasons for an action are considerations which count in favour of that action," denies (2013, 46ff.) the primacy of reasons

someone does something, a reason that, in the agent's eyes, counts in favour of her acting in a certain way" (Alvarez 2016a). This distinction, largely present in different authors (see, e.g., Dancy 2000; Raz 1999a; Parfit 2011), can be further articulated considering that motivating reasons can play a double role: reasons that motivate someone to act can also be one that, in a third-person role, could also be used to explain a behaviour. In this way, it is possible to arrive at a "three-part classification of reasons:" *normative*, *motivating*, *explanatory* (Alvarez 2016a). The need to distinguish between motivating and explanatory reasons, despite the fact that "one might think, fundamentally," that they are "the same," can be seen in the fact that "even if the same reason sometimes answers the two questions about motivation and explanation, this is not always so" (ibid.). Although a reason that motivates an action can always explain it, a reason that can explain the action is not always the reason that motivates it" (ibid.).[11] From this perspective, as Audi (2010, 273ff.) remarks, we can individuate "three overlapping kinds" of reasons: "normative reasons—which include moral reasons as a major subset—motivational reasons, and explanatory reasons": "Normative reasons are reasons (in the sense of objective grounds) there are for doing something"; motivational reasons are "reasons someone has to do something"; explanatory reasons "are reasons why someone acts."[12] Following a suggestion by Alvarez (2016a), referring to Dancy (2000, 2ff.), this distinction between reasons can be seen as the way in which we can answer the different questions that reasons pose in relation to action and reasoning: in this way, normative reasons will be those that answer the question of "whether there was good reason to act in that way […], *any reason for doing it* at all"; motivating reasons will be those that answer questions "about *his reasons for doing it*" (Dancy 2000, 2); and explanatory reasons will be those that answer a general "reason why" different "sorts of things"—like "the occurrence (or nonoccurrence) of an event; the

in normativity and supports that of the "ought." To pro tanto reasons, he adds the presence of "pro toto reasons" (or "perfect reasons"). He does not "take the idea of a reason as primitive" (ibid., 54), and he does not consider "counting in favour [as] the basic normative notion" (Broome 2004, 41). For Broome, the idea of a pro tanto reason does not account for the explanatory role of reasons. On these aspects, see Crisp (2014). See also note 35 below.

[11] For Alvarez (2016a), "the advantages of drawing this distinction will be spelled out in examining debates concerning motivating reasons and the explanation of action. […] [A]pparently competing claims about motivating reasons and the explanation of action are often best understood and resolved as claims about motivating or explanatory reasons, respectively. To be precise, a reason that plays a motivating role for a particular action can (arguably) always play an explanatory role for that action, but the converse does not hold." She argues that "This way of categorising reasons […] enables us to deal with a range of cases that the binary classification cannot accommodate" (ibid.). To explain the difference between the two types of reasons, she uses the example of Othello and Desdemona: while Othello's jealousy can be seen as one of the reasons explaining the killing of Desdemona, the reason motivating Othello is the belief, induced by Iago, that Desdemona betrayed him (ibid.).

[12] Audi (2010, 275) notes that normative reasons can be, for example, moral or prudential. He also notes that "some normative reasons for—roughly, counting in favour of—an action are reasons for any normal human being. Other normative reasons, however, are person-specific: reasons there are for a specific person."

obtaining (or non-obtaining) of a state of affairs; someone's or something's φ-ing (or not φ-ing)"—happened or not (Alvarez 2010, 27).[13]

This distinction between normative (reasons that favour), motivating (reasons for which we act), and explaining (reasons why) should be developed not only to clarify the exact role of reasons but also to explore what it means that reasons favour, motivate, and explain.

2.1 Normative Reasons[14]

2.1.1 Normative Reasons and the Source of Normativity

The idea that a reason is normative refers to a general notion of what normative is: *normative* means "prescriptive" in relation "to some norm or value and, by implication, concerning correctness," that is what makes something "right or appropriate" in relation to what is prescribed by "norm or values."[15] Normative is therefore what we can consider "right or wrong with reference to what is prescribed by the […] norm, or what furthers the […] value" (Alvarez 2010, 9),[16] and in this sense, normativity implies the correctness of actions in relation to norms or values.

For a reason to be normative, it must therefore "involve the idea that reasons can be invoked to support claims about what it would be right (for someone) to do, believe, want, feel, etc." Normative reasons are those that "can be invoked to support claims about what it would be right (for someone) to do, believe, want, feel, etc." (ibid.),[17]

---

[13] Broome (2004, 34) underlines that "a useful distinguishing mark is that 'the reason' in this non-normative sense is usually followed by 'why.'" For Alvarez (2016a), "the basis for doing so was said to be the existence of three distinct questions about reasons: whether a reason *favours* an action; whether a reason *motivates* an agent; and whether a reason *explains* an agent's action. Accordingly [...] we should recognize three kinds of reasons: normative, motivating and explanatory."

[14] Normative reasons are often qualified "in terms of justification: a reason justifies or makes it right for someone to act in a certain way. this is why normative reasons are also called 'justifying' reasons" (Alvarez 2016a). Dancy (2000, 6–7) criticizes this way of marking normative as justifying reasons: "there is a sense of 'justify' in which I can be said to have justified doing what I did. but this does not show that the balance of reasons was in favour of the action. It wasn't. […] So I prefer to keep the notion of a reason that justifies separate from that of a normative reason. Broome (2004, 54), too, notes that "'justify' is an ambiguous word."

[15] Norms and values are understood here in a broad sense, that is, as inclusive of different values (moral, prudential, etc.) and norms (legal, social, principles, codes, etc.). For Alvarez (2016a), "the existence of these norms or values depends on a variety of things": "logical and natural relations, conventions, rules and regulations, etc."

[16] From the same perspective, Alvarez (2016a) notes that "'normative reason' derives from the idea that there are norms, principles or codes that prescribe actions: they make it right or wrong to do certain things." Dancy (2000, 1) sees normative reasons as "good reasons for doing the action. So they are *normative,* both in their own nature (they *favour* action, and they do it more or less strongly) and in their product, since they make actions right or wrong, sensible or unwise."

[17] Some authors note that the idea of favouring is an ambiguous one. Schroeder (2007, 11), for example, notes that the idea that reasons "count in favour […] is slippery." Hieronymi (2005,

or, in other words, they are "a reason [that] justifies or makes it right for someone to act in a certain way" (Alvarez 2016a). In this way, that a "reason can favour φ-ing" means that "it can make φ-ing right or appropriate:" to favour is to contribute to or support (i.e. to "recommend, warrant, demand") the rightness or appropriateness of an action (Alvarez 2010, 3, 11). Making right or appropriate a certain action can depend on many factors, such as values, norms, codes, rules, desires, purposes, and natural facts.[18] In addition, norms and values especially may refer to different fields and may be "moral, prudential, legal, hedonic (relating to pleasure) or of some other kind" (Alvarez 2016a). Finally, the relevance of these factors may radically change depending on the different contexts.[19]

The normative aspect of reasons is expressed "by talk of 'a reason *for* φ-ing' (or 'a reason *to* φ')": in this sense, we have reasons for acting, believing, deliberating, wanting something, feeling an emotion, and so on,[20] and, as noted, "there are reasons for the variety of things where we are, typically, responsive to reasons."[21] Having a reason for φ-ing is often expressed through the concept of "ought": saying "that if there is a reason for someone to φ" could be seen as saying "that person […] *ought* to φ." This does not mean having a moral ought, nor does it mean that the behaviour favoured by a reason is "obligatory": what "ought to" amounts to varies from case to case, depending on the circumstances (Alvarez 2010, 11).[22]

437–438, 456–457) makes a deeper critique, stating that "we should not understand 'counting in favour of an action or attitude' as the fundamental relation in which a consideration becomes a reason." She stresses that identifying a "reason as a consideration that counts in favour of an action or attitude […] generates a fairly deep and recalcitrant ambiguity; this account fails to distinguish between two quite different sets of considerations that count in favour of certain attitudes, only one of which is the 'proper' or 'appropriate' kind of reason for them." This ambiguity, which for Hieronomy, leads to the so-called problem of the "wrong kind of reason," referring to the fact that seeing a reason as "a consideration that counts in favour of an action or attitude […] generates a thoroughgoing ambiguity in reasons for certain attitudes–we cannot distinguish precisely between 'content-related' and 'attitude-related' reasons." Instead of seeing a reason as counting in favour of something, she proposes that we see it "as a consideration that bears on a question […] in a piece of reasoning." This makes it possible to "distinguish between questions, thus distinguishing these classes. The attitude-related reasons count in favour of the attitude by bearing on whether the attitude is in some way good to have; the content-related reasons count in favour of the attitude by bearing on some other question." On this analysis, see also Hieronymi (2013).

[18] It is possible to relate these factors to values and norms understood as reference points for desires and purposes, among other things. However, this relation need not be necessary. In any case, we will refer to values and norms as categories in relation to which something can be considered as having been made "right or appropriate."

[19] As Alvarez (2010, 11) notes: "Thus, if A ought to φ, the circumstances of each case will determine whether A's φ-ing is merely recommended, or whether it is also required, or mandatory." On the way in which it is possible to consider the role of context, see Sect. 5.1 below.

[20] Or, conversely, "reasons for not φ-ing, that is reasons for not doing or for not believing something, for not feeling something, etc." (Alvarez 2010, 10).

[21] Alvarez (2010, 10), referring to Raz (1999b).

[22] This is true for the positions that consider reasons as the fundamental dimension of normativity. As we have seen, this is not, for example, the position of J. Broome.

This definition (normative reasons as what favour *φ-ing*) can be interpreted in different ways in connection with what can be seen as the source (ground, basis, capacity)[23] of the normativity of reasons, that is, of their ability to favour action. Schematizing and simplifying, we can individuate three main positions: the first *desire-based*, the second *value-based*, and the third *rationality-based*. On the first position, "all reasons for acting, intending, and desiring are provided by the fact that the agent wants something or would want it under certain conditions" (Chang 2004, 56). In this perspective, "all practical reasons are grounded in the present desires of the agent; justification has its source in the fact that I do or would want it" (ibid.).[24] In this sense, desires can be seen as "a necessary condition for a consideration to provide an agent with a reason" (Sobel and Wall 2009, 3). On the second position, "no practical reasons are provided by the fact that one desires something." In "'value-based' accounts, reasons for acting, intending, and desiring are provided by facts about the value of something, where being valuable is not simply a matter of being desired" (ibid., 57).[25] There are two principal versions of this second approach: the "buckpassing versions" and the one linked to idea of the existence of evaluative facts. On the buckpassing version, it is not strictly the evaluative fact that provides a reason but the facts on which the evaluative fact supervenes, while on the second, reasons come from evaluative facts. In this way, "value-based views ground all practical reasons in evaluative facts or the facts that subvene them" (Chang 2004, 57),[26] and "justification has its source not in the fact that one wants something but in facts about what one wants" (ibid.).[27] The third (rationality-based) approach is Kantian, and at the centre of normativity, it places the idea of an autonomous and rational subject: "Kantians […] hold that rationality is the source of practical normativity." From this perspective, "rational deliberation legislates […] what to do and an action's being the rational thing to do is what makes it true you ought to do it." As it is well know, "Kantians hold that there are non-instrumental rational deliberative procedures guaranteed to issue true normative conclusions," that is "principles

[23] Different authors mention this aspect in various ways. For example, Robertson (2009, 18) speaks of "source," Alvarez (2016a) of "basis" and "capacity," Sobel and Wall (2009, 3) of "grounds."

[24] Chang (2004, 56) explains this approach in this way: "My reason for going to the store, for example, is provided by the fact that I want to buy some ice cream, and my reason for wanting to buy some ice cream is provided by the fact that I want to eat some." This is a form of "subjectivism," that is a view that assumes the capacity of reasons to favour desires, plans, motivations, projects, etc.

[25] Note that, in this case, "my reason to go to the store is provided by the value of what is in question—namely, eating some ice cream—and the value of eating some ice cream is given by the fact that doing so would be valuable in some way—for example, that it would be pleasurable. It is not the fact that I want ice cream that makes having some pleasurable; having ice cream might be pleasurable even if I don't desire it" (Chang 2004, 56).

[26] Chang (2004, 57) describes the difference in this way: "So, for example, my reason to have the ice cream might strictly be given not by the evaluative fact that it would be pleasurable but rather by the natural facts upon which its being pleasurable supervenes, such as that it would be pleasant or that I would enjoy it."

[27] One of the aspects in which these approaches diverge is that of the internal versus external vision of reasons for acting. On this question, see Sect. 3 below.

or laws that any rational agent could recognize and that thereby apply to any rational being" (Robertson 2009, 11, 19). As noted, "on the Kantian view, ideal rationality significantly constrains what is desired or willed in the authoritative way" (Sobel and Wall 2009, 4).

As we will see, these different approaches (and in particular the desire- and value-based ones)[28] can have different implications with regard to the different roles of reasons: in discussing the normative aspect, one of these aspects is the relation to the concept of rationality (understood as the way to arrive at normative conclusions). Very briefly, it is possible to note that the desire-based approach underlines the relevance of an instrumental (means-ends) rationality and the requisite of the "connections between an agent's various attitudes"[29]; Kantian approaches emphasize a procedural form of rationality[30]; the value-based approach stresses a "substantive conception of rationality" grounded in the idea that "there are substantive normative truths" (Robertson 2009, 20).

2.1.2 The Structure of Normative Reasons

A normative reason has a "relational" structure: "it establishes a relation between a fact, an agent, and an action kind" (Alvarez 2016a). As has been noted, "the concept of a reason is itself relational. Basic reason statements of the form 'A has a reason to φ' or 'there is a reason for A to φ' indicate a relation holding between some agent A and some act φ (e.g. an action, belief, feeling)—a reason is a reason *for* someone and *to* or *for* something" (Robertson 2009, 9).[31]

---

[28] The Kantian approach can be assimilated to a "subjective" one characterized by the control that the rationality of an "ideally rational agent" imposes on desires and attitudes. See Sobel and Wall (2009, 3ff.).

[29] A model of practical rationality that refers to mental states has been developed in Bratman's theory of action (1987, 2014): this is the belief–desire–intention (BDI) model. This model has as its main reference the concept of plans for action and has been used in the field of artificial intelligence research.

[30] Robertson (2009, 19) states that "Kantian principles of rationality are formal in that they lack specific, substantive, normative content, a practically rational deliberator being one who satisfies relevant formal procedures of reasoning."

[31] Robertson (2009, 12), sees it as "uncontroversial […] that the conceptual structure or logical form of a reason is that of a relation." Raz (1999a, 19) notes that "we usually think of reasons for action as being reasons for a person to perform an action." Blackburn (2010, 6) underlines that "reasons are reasons for something: the primary datum is relational. The field of the relation is less clear, or rather, more diffuse." For Searle (2001, 99), "reason statements are relational in three ways. First, the reason specified is a reason for something else. Nothing is a reason just by itself. Second, reasons for action are doubly relational in that they are reasons for an agent-self to perform an action; and third, if they are to function in deliberation, the reasons must be known to the agent-self. To summarize, to function in deliberation a reason must be for a type of action, it must be for the agent, and it must be known to the agent." Alvarez (2016a) stresses that, for some authors like Skorupski (2010) and Scanlon (2014), "the relation involves not just a person, a reason and an action, but more aspects: a time, circumstances, etc."

What determines the way "in which a reason makes φ-ing right, and hence in which something may be right or justified," depends not only on the different factors (like types of norms and values) but also on "what φ-ing is" (Alvarez 2010, 13). We have significant differences between the case of believing, that of acting, and that of feeling emotions. In the first case (believing), "the rightness or appropriateness of φ-ing […] concerns the concept of truth"; in the second (acting and wanting), "it concerns the concepts of what is valuable and of the good"; in the third (emotions), making right has to do with the appropriateness/reasonableness of feelings or emotions, that is with their being "fitting or proportionate to the facts" (ibid.). On this basis, it is possible to distinguish, within normative reasons, among epistemic reasons related to the concept of truth; those that are practical, relating to action and deliberation; and those related to "emotions."[32] The distinction between epistemic and practical reasons (for action) is particularly relevant because, as we shall see in Sect. 4, the former do not imply a choice between different values and do not seem "person-relative," while the latter do imply a choice between different values and *are* "person-relative."[33] As we shall see, this distinction is related to the fact that epistemic reasons have a single (or prevalent) reference criterion (the truth), while practical reasons refer to a "variety of values."

Normative reasons are in general defeasible, that is, that they can be "defeated by a reason for not φ-ing" (Alvarez 2010, 12). This means that normative reasons are "pro tanto" (or prima facie) reasons. This indicates that a reason "can" favour an action (belief, act, evaluation, etc.) to a certain extent, but that this possibility is limited by the presence of reasons which instead favour an omission. So "I have a *pro-tanto* reason" to do something "and a different *pro-tanto* reason" not to do it (Alvarez 2016a).[34] As Broome (2004, 41) has suggested, a pro tanto reason refers to "a weighing explanation" and to "the distinction between the for-φ role and the against-φ role in a weighing explanation."[35] There is, however, no agreement that all reasons are pro tanto: the same author argues that, next to pro tanto reasons, there

[32] We will not discuss these reasons. Raz (2009, 47ff.) sees "reasons for or against having an emotion" both as standard (adaptive) reasons ("affect-justifying reasons") and as non-standard reasons. On this distinction, see Sect. 4 below.

[33] The notion of person relatedness can be connected with the distinction between agent-relative and agent-neutral reasons: it is a distinction that was introduced by Parfit (1984) on the basis of the distinction between "objective" and "subjective" reasons elaborated by Nagel (1970). For Parfit (1984, 142), "Nagel calls a reason *objective* if it is not tied down to any point of view. Suppose we claim that there is a reason to relieve some person's suffering. This reason is objective if it is a reason for everyone—for anyone who could relieve this person's suffering. I call such reasons agent-neutral. Nagel's *subjective* reasons are reasons only for the agent. I call these agent-relative […]. When I call some reason agent-relative, I am not claiming that this reason *cannot* be a reason for other agents. All that I am claiming is that it may not be." On the different ways of understanding this distinction, see Ridge (2011).

[34] Dancy (2000, 5) argues that "there can be good reasons not to do an action even when there are better reasons to do it. That an action was right does not show that there were no (good) reasons not to do it. Equally, if an action was wrong, this does not mean that there was no reason to do it; it merely means that there was insufficient reason."

[35] Broome (2004, 42ff.) is against what he calls "protantism," that is the view that "there is a case for thinking that every ought fact has a weighing explanation." For Broome, "even if every ought fact

are "perfect reasons to φ," in which "to φ is defined as a fact that explains why you ought to φ," that is when there is a direct link between a non-normative (explanatory) fact and the proposition that someone ought to φ. Broome differentiates these two kinds of reasons as follows:

> A perfect reason for you to φ is a fact that explains why you ought to φ. A pro-tanto reason for you to φ is a fact that plays a characteristic role in a potential or actual weighing explanation of why you ought to φ, or of why you ought not to φ, or of why it is not the case that you ought to φ and not the case that you ought not to φ. (Broome 2004, 55)[36]

The same author, however, seems to emphasize that whether a reason is conceived as "perfect" or as "pro tanto" depends on a specific philosophical position (evidentialism vs. pragmatism): this means that it is possible to consider reasons in the two ways. In addition, one can add that a perfect reason can be a reason that does not have a "current" reason for not φ-ing, but that this reason can still be hypothesized. If we hold the hypothesis that normative reasons are, in general, pro tanto reasons, an important consequence arises, that is as we shall see in Sect. 5, that in a process of weighing, reasons will have different strengths, roles, and interrelations; this means that the "final" reason for doing φ (acting, believing, deliberating, etc.) should be considered as an overall, or "all things considered," reason to φ, that is a reason that "overrides" or "defeats" other competing reason(s).

2.2 Motivating Reasons

A "motivating reason"[37] is "a reason for which someone does something," that is "a reason that, in the agent's eyes, counts in favour of her acting in a certain way" (Alvarez 2016a). In greater detail, it can be seen as a reason that an agent "took to make his *φ*-ing right and hence to speak in favour of his *φ*-ing, and which played a role in his deciding to *φ*" (Alvarez 2010, 35).[38] In other words, it is "a reason that the agent takes to favour her action, and in the light of which she acts" (Alvarez 2016a). A motivating reason, that is a reason that can be recognized "when an agent acts

---

does have a weighing explanation, many ought facts also have more significant explanations that are not weighing ones." As noted, Broome denies that reasons are the primary category of normativity. In his view, this space is composed of different types of reasons (perfect, or pro toto, and pro tanto) and normative requirements. The "weighing conception" of reasons is also questioned by Horty (2007, 1–2) who proposes to conceive reasons as "defaults:" he sees the weighing conception as "incomplete as an account of the way in which reasons support conclusions."

[36] For Broome (2004, 42–43), "a putative example of an ought fact that has no weighing explanation" is that "You ought not to believe both that it is Sunday and that it is Wednesday."

[37] Alvarez (2010, 54) stresses that "the term 'motivating reason' does not have much currency outside of philosophy and can rightly be regarded as a term of art."

[38] Audi (2010, 275) distinguishes between "motivational" reason from a "motivating" one: the former is a "potential" reason "even if I never act on it," while the latter is one that "explains why I do" something. Motivational reasons must be "*possessed* reasons: reasons someone *has* to do something."

[…] in light of that reason," works as a premise "in the agent's (implicit or explicit) reasoning about φ-ing," that is as "a premise in the practical reasoning […] that leads to the action" (Alvarez 2010, 35). It should also be noted that speaking of a singular reason "of an agent's motivating reason" or of "the agent's reason" is, of course, a "simplification," both because "an agent may be motivated to act by more than one reason," and because "a fact will seem a reason for me to act only in combination with other facts" (Alvarez 2016a).[39]

This definition leaves some issues open in relation to what a motivating reason is and, consequently, what the elements that constitute it are. Two are the main problems: on the one hand, the question of the role that the subject's desires, goals, and willing have in motivating action and, on the other hand, the role of the beliefs. In the first case, the problem relates to the ability that the reasons have to motivate agents to act (or not act): this is a matter of understanding the transition from *having reasons* to φ-ing to *doing* φ. We have to explain "how thinking that there is a reason for me to do something can motivate me to act, and to act *for* that reason" (Alvarez 2016a) and so to answer the question "about the conditions that determine when a reason for acting applies to a particular agent" (ibid.). In the second case, the problem is that of the role that knowledge and beliefs have in motivating an action: whether they are requisites of motivation and can be considered as motivating reasons. This seems particularly relevant in the case of "false belief," that is when it seems that an agent acts on the basis of a untrue belief.[40]

In relation to these questions it is possible to identify two positions: the first, related to the desire-based approach (and more generally to a Humean view of reasons), is that of "psychologism," while the second is that of "non-psychologism" or "factualism" (Alvarez 2016b). According to the first position, the answer to the two questions implies the central role of desires and beliefs (even disjointly), while according to the second, desires and beliefs can only be seen as aspects that generally motivate, but not as specific reasons for acting. According to the first position, desires and beliefs have a central role, "whether the reasons that apply to you depend on your desires and motivations," and at the same time, "for a reason to motivate you it must be a reason you have," requiring that you "possess" (Audi 2010, 275) the reason, and therefore, that you "must know or believe the consideration that constitutes the reason" (a mental state) (Alvarez 2016a); the second position, by contrast, denies the role of desires and of beliefs.

According to the first position, "it seems right that when an agent acts for a reason, he acts motivated by an end that he desires (an end towards which he has a 'pro-attitude') and guided by a belief about how to achieve that end" (Alvarez 2016a). According to the second position, the role of desires and beliefs is instead not primary, and motivating reasons are facts. This position distinguishes between

[39] Alvarez (2010, 127) notes that reasons for φ-ing can be either "independent" of or "related" to one another. The criterion for the distinction "is the relation between reasons for acting and the goodness or value of the action for which I take them to be reasons that explains why sometimes my reasons for doing something are independent of each other and why sometimes they are not."

[40] As we have already noted, Alvarez (2016a) highlights this problem using the example of Shakespeare's *Othello*.

"being motivated" (inclined) to do something and having a "motivating reason": desires, motivations, willing, purposes, and so on can be seen as things that motivate, but not as real motivating reasons (Alvarez 2010, 53ff.).

These two positions refer to different views about the ontology of the reasons. As we will see in Sect. 3, this is the question of the "conceptual category or categories" to which reasons belong: the category of mental states or that of facts (Alvarez 2010, 32). What must be noted is that these two positions include or exclude different aspects according to ontological choice. Thus, on the first vision, they will be "someone's goal or intention in acting, which is something that the agent desires" (Alvarez 2016a), plus an actor's belief, while on the second "they do not fall under the category 'motivating reasons'" and they can at most be states "that encompass motivation" (ibid., referring to Mele 2003) and thus general motivation, "not motivating reasons" (ibid.).[41]

2.3 *Explanatory Reasons*

We can define *explanatory reasons* as the "reasons why someone acts" (Audi 2010, 275). This means that the "explanans is 'the reason why' someone acted" (Alvarez 2010, 161).[42] A reason explanation is the "reason in the light of which [a subject] φ-ed" (ibid., 36).[43]

There are various kinds of action explanation[44]: following a suggestion of Alvarez, we can distinguish, with reference to the explanantia, three principal groups of explanatory reasons: "explanations of action that take the form 'A φ-ed because q', where 'q' is (i) the reason for which the agent acted […]; (ii) a reason why A φ-ed which is a fact concerning A's beliefs and desires, knowledge, emotions, feelings, motives, character traits, habits, etc.;" (iii) a reason that explain "by citing a purpose or goal that the agent pursued in his action, and they are formally characterized by the use of 'in order to', 'with the purpose of', 'for the sake of', or equivalent expressions" (ibid., 191).

The first ones can be defined as "reason explanations *proper*," the second as "*psychological explanations*," the third as *purposive* (*teleological*, *intentional*) *explanations*. In the first case, we cite in the explanans a reason of the agent that may also

[41] In Othello's case, according to the first position, the reason for killing Desdemona lies in his desire "to restore his reputation," crippled by Desdemona betrayal, while according to the second, the reason lies in "the putative facts that she is unfaithful to him and that killing her is a fitting way to restore his reputation." For a detailed analysis, see Alvarez (2016a).

[42] Alvarez (2010, 28) notes that "in addition to answers to 'why?'-questions, there are other kinds of explanation, for instance, explanations of how to φ (how to iron a shirt), of how x φ-s (how a steam engine works), etc. But to explain how to φ, or how x φ-s, is not to give a reason."

[43] Alvarez (2010, 166) stresses that the explanation is about "why someone habitually acts, is now acting, has acted, will act, etc., in a certain way," making these actions "intelligible."

[44] The answers that we can give depend also by "the pragmatics of explanation," that is "the context in which the question is asked" (ibid., 26).

be the "reason for which the agent acted" (ibid., 5): so we have a proper explanation because we have an overlap of explanatory and motivating reason.[45] In the second case, the explanans refers to "a psychological fact about the agent, such as the fact that he believed and wanted certain things, or the fact that he had certain motives, or character traits, emotions, habits, and so on" (ibid.). This second sort of explanation can be called a "'Humean explanation', which typically has the form 'he φ-d because he believed that p and wanted x'" (ibid.). In the third case, "actions performed for a reason can be explained with purposive or intentional explanations," that is "explanations of actions [that] identify the agent's purpose, which is also normally his intention in acting." These "are a kind of teleological explanation" of the form of "'he φ-ed *in order to* ψ'" and "characterized by the use of 'in order to' or equivalent expressions ('so as to', 'with a view to', etc.)" (ibid., 171).

As is the case with motivating reasons, these different kinds of explanations find their foundation in a different ontological view of reasons. This fact is reflected not only in the difference between non-psychologistic explanation (which invokes a fact) and psychologistic explanation, but also in the different relevance that these views attribute to purposive explanation and to the distinction between reasons for action and causes.

Unlike psychologistic approaches, non-psychologistic ones see "purposive" explanations as nonindependent ones that must be "supplemented by reason explanations." In fact, "these explanations […] do not *state* the fact that the agent had such a goal. And, when they are explanations of things done for reasons, they do not state the reason the agent had for doing what he did, though they often suggest what that reason was" (ibid., 170). On this approach, explanations that refer to the agent's intentions make it necessary to further identify the reasons for the action, as the purpose and intention are merely the general directions of the action: "the goal or purpose is the end towards which the action is directed," and the reason is "a fact that guided the agent in his pursuit of the goal mentioned in the explanation" (ibid., 194).[46]

Another distinction relates to the relationship between reasons to act and causes. Of course, it is possible to distinguish between causal explanation and reason explanation on the basis of the fact that the former is a "natural" relationship between facts (unintentional) and the second concerns rational actions. However, what distinguishes psychologistic explanations from non-psychologistic ones is that the former would admit of a causal explanation of the action. As we shall see, this point makes it possible to claim that psychologist explanations are more appropriate: only they can be a premise in a causal explanation. On a non-psychologistic approach, a

[45] For Audi (2010, 276), "they are reasons for which we do something and thereby ground a motivational explanation of our doing it."

[46] You can have it as your purpose to be on time for a meeting but have different reasons why you are taking a taxi (someone stole your bicycle, you woke up late, the bus is late, and so on). Of course, there are actions that can be explained on the basis of their purpose: they are "merely reactive" (instinctive, mechanical, reactive, habitual, etc.) and "skilled actions" (riding a bike, skiing, dancing, driving, etc.) that "can be explained without the need to attribute to the agent any explicit or implicit calculation or reasoning" (Alvarez 2010, 193, 195).

reason-based explanation does not make it necessary to identify reasons as causes: in explaining an action, it is therefore important to identify the reasons "regardless of whether the explanations in which they feature are causal or not" (ibid., 30).

3 The Ontology of Reasons

Ontological reflection addresses the question of what is a reason, or what conceptual category it is possible to assign it to. As we have already seen in relation to the different roles of reasons, there are different characterizations of how to understand these roles and their actual meaning. So the fact that reasons can play different roles can itself be taken as a clue to their ontological diversity.

There are two main positions: on the one hand, psychologistic ones, which emphasize the role of beliefs and desires, and which consider reasons as mental states, and on the other hand, non-psychologistic (or factualist) ones, which see the reasons as facts.[47] Psychologistic positions assign a double ontological nature to reasons—that of facts and of mental states—while non-psychologistic ones identify only a single ontological determination, that of facts. Prevalent in contemporary literature is the idea that these two positions are in accord in relation to normative reasons, viewed in general as facts, but diverge in relation to motivating and explanatory reasons.[48] Although the divergences mainly concern these two aspects, they also affect the idea of normativity and, in particular, of their normative capacity. We will therefore discuss the different visions in general terms.

What is relevant is not just the subject of discussion (reasons as facts or as facts and mental states) but also, and especially, the implications these two positions have for the analysis of the reasons. In a vast debate, two crucial problems can be identified: the first is to determine what normative and motivating reasons are; i.e., whether or not in order to have a reason or for it to motivate, it must be part of an agent's mental state. The second concerns the role of beliefs: in this case, it is not just a matter of assessing not only the role of mental states, but also the rationality of action. If practical reasoning implies the presence of premises (from which to draw motives, decisions, conclusions, etc.), it seems necessary that these at least be known (and believed) by the agent. These two positions refer to widely debated issues in ethical and metaethical reflection. In particular, the issue of motivation refers, inter alia, to the contrast between internalism and externalism, while the problem of beliefs refers

[47] Alvarez (2016b) identifies different kinds of non-psychologism (factualism). Kantian perspectives are difficult to locate but, as just noted, they seem to be close to subjectivism (psychologism). For Scanlon (2014, 11), Kantians start from "the idea that claims about the reasons an agent has must be grounded in something that is already true of that agent," and this is "something similar might be said by proponents of desire-based."

[48] Alvarez (2016b) asserts that "Factualism is widely accepted for normative reasons, reasons that favour doing something. But things become more controversial when the question concerns the reasons for which we act and the reasons that explain our actions. This has led many to conclude that Factualism is right for normative reasons but not for other kinds of reasons."

to the distinction between objective and subjective (on the model of the distinction between objectivism and perspectivism). Further, distinctions are then present in the various settings: particularly, relevant is the one present in non-psychologistic positions that, in relation to the facts, separates "realism" and "irrealism."

The psychologist position can be seen as the prevailing, and in some ways "orthodox," one, while the non-psychologistic one develops from criticism of the former (Alvarez 2010, 2). This first position was made canonical by Davidson (1963) in an essay titled "Actions, Reasons, Causes," in which he describes the explanation of an action through a reason as rationalization. Davidson identifies what he calls a primary reason. In a well-known passage, he writes:

> Whenever someone does something for a reason, therefore, he can be characterized as (a) having some sort of pro-attitude[49] towards actions of a certain kind and (b) believing (or knowing, perceiving, noticing, remembering) that his action is of that kind." [Consequently] giving the reason why an agent did something is often a matter of naming the pro-attitude (a) or the related belief (b) or both; let me call this pair the primary reason why the agent performed the action. (Ibid., 685–686)

To this consideration of what a primary reason is, Davidson adds that only under these conditions can a reason be part of practical reasoning (which for Davidson is causal reasoning).[50]

As is apparent, in this reconstruction reasons refer to mental states such as wishes and beliefs. As noted, this seems to have important advantages with regard to both the motivation of action and the rationality of action. In the first case, the fact of being dependent on a subject's mental states makes it comprehensible because a normative reason applies to a subject. As noted, "desire-based accounts of reasons may seem to have the edge here" because the idea that a reason determines my behaviour is far more persuasive if it "depends on my antecedent motivations (desires, plans) and therefore that a subject is motivated what he believes" (Alvarez 2016a).

As noted, this issue, from a metaethical perspective, is marked by the distinction between internalism and externalism: the advantage of the first depends on the fact

---

[49] For Davidson (1963, 685–686), "under (a) are to be included desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values in so far as these can be interpreted as attitudes of an agent directed towards actions of a certain kind. The word 'attitude' does yeoman service here, for it must cover not only permanent character traits that show themselves in a lifetime of behaviour, like love of children or a taste for loud company, but also the most passing fancy that prompts a unique action."

[50] Buckareff (2014) stresses that Davidson "argues that the relationship between reasons and actions displays the same pattern we discern in causal explanations" and that "if the onset of a primary reason is not a cause of action, we have difficulty accounting for the difference between when an agent has a reason for acting in mind that does not actually explain why she acts as she does and cases where the agent has a reason for acting that does explain why she acts as she does. If we dispense with a causal role for reasons, we may be able to appeal to some reasons of an agent in the light of which the action looks reasonable, but, absent a causal role, it is not clear that the putative justification for action explains the action and, hence, *really* rationalizes it." O'Connor (2010, 130) notes that in the perspective of the "casual theorist," it is possible "determining the true reason(s) for the action."

that the reference to pro-attitudes makes the motivation immediate and requires no additional factors (double pass). In fact, as a general scheme, it is argued from this perspective that "every reason for action must bear relation *R* to motivational fact *M*" (Finlay and Schroeder 2012), while in externalism this relationship is considered unnecessary.[51]

The second aspect that favours the psychologistic approach has to do with the role of beliefs. There are two main features: the first, as mentioned, is linked to the need to identify the reasons for an agent's behaviour requires the agent's "believing or knowing": "a fact that is merely 'out there' cannot explain why you do anything," and therefore the reasons for acting are "mental states (believings, knowings, etc.)" (Alvarez 2016a). The second feature has to do with the ability of the psychologistic approach to explain the action that takes place on the basis of "false beliefs" ("error case"). The fact that actions can be made on the basis of such beliefs seems explicable only from the central role of mental states. Without this acquisition it would also seem impossible to explain the rationality of action (in the minimal sense of drawing conclusions from premises). In fact, some actions take as their starting point a false (albeit sometimes reasonable) belief that justifies and explains them. Reasons in this sense are to be understood in the light of the agent's subjective perspective: psychologistic theories therefore support a "perspectivist" position, as the fact of having a (normative) reason "is not independent of [the agent's] perspective, which includes her beliefs" (ibid.).

Non-psychologistic positions, as mentioned, claim that the reasons (especially normative ones) are facts. They question the main assumptions of the psychologistic theses and point out their conceptual ambiguity. As to the relevance of desires, these positions, on the one hand, emphasize that action often does not occur on the basis of desires and that this can be detected in relation to moral norms and to a wide range of cases, which include, e.g. commands, orders, while on the other hand they distinguish between "being motivated" and reasons to act: from this perspective, desires, beliefs, purposes, pro-attitudes, etc., can be seen as "inclinations," but such inclinations are put into effect on the basis of reasons understood as facts (see Alvarez (2010), 53ff.).[52]

More complex seems the reply to problems posed by false beliefs and cases of error. In this context, the answer that appears most convincing is that reasons are

[51] Finlay and Schroeder (2012) refer this definition to schematic internalism and stress that "different ways of spelling out relation *R* and motivational fact *M* correspond to […] a different thesis—a *version* of reasons internalism." Darwall (1992) distinguishes between "existence" and "judgment" internalism: the latter is based on the idea that "assent to a moral judgment […] concerning what one should do is necessarily connected to motivation (actual or dispositional)"; according to the first, "someone morally […] has to do something only if, necessarily, she (the agent) has (actually or dispositionally) motives to do so."

[52] The scheme of this analysis can be summarized as follows: the motivations (which cannot be reduced exclusively to dispositions, as in Ryle's analysis) are mainly understood as desires ("to give the motive for a deed is to indicate a desire for the satisfaction of which the deed was done"), and a desire is in turn seen as an "inclination to act." See Alvarez (2010, 60, 61, 70), referring to Ryle (1949) and White (1968) (quotation in parentheses) and criticizing Smith (1987) and his attempt to characterize desires through the use of the metaphor of "direction of fit."

"apparent reasons," that is "something [that] appeared to the agent to be a reason but it was not really one" (ibid., 140). This answer is substantiated by highlighting what appear to be the main ambiguities of psychologistic theories, namely the possibility of interpreting mental states either as acts or as the content of acts (the "act/object ambiguity"): "the term 'belief' [and also 'desire'] can be used to refer to one's believing something or to what one believes" (the content of a belief) (ibid., 125). If we use the term *belief* in the second sense, "to say that reasons are beliefs is not to say that they are mental states, because, even if believing that p is a mental state, what is believed (that p) is not itself a mental state" (ibid., 45). The two visions of what a belief is "are quite different": "reasons might be beliefs and desires and yet not be 'believings' and 'desirings' and therefore may not be mental states" (ibid.). Considering beliefs (and desires) as what is believed means referring "to a proposition," and this signifies that if "we think of […] reasons as what is believed, as facts, as propositions, etc., it is clear that what we are thus thinking about are not psychological entities" and we are not committed "to the view that motivating reasons are psychological entities."[53] In sum, "what is believed is not a psychological entity" (ibid., 154). In relation to error cases, then, this approach "does not lead to any implausible conclusions […] because in such cases the agent acts only for an apparent reason—and the apparent reason he acts for is not that he believes that p, but rather what he believes, namely that p" (ibid., 146).

What kinds of facts do the non-psychologistic theses refer to? They can be said to generally use a broad concept of fact. The idea that "all reasons are facts or […] truths, which are expressed propositionally, and which can be premises in reasoning, both theoretical and practical" is declined in a "minimalist" conception (as opposed to an "austere" one) that makes it possible to include the greatest number of facts (including, e.g., negative facts and values).[54] Such a conception can be stated either by saying that a "fact is a true proposition" (Alvarez 2016a) or, in a more undemanding version, by affirming that by "'fact' is meant simply that which can be designated by the use of the operator 'the fact that…'" (Raz 1999a, 17–18).[55] Two aspects are finally to be considered: first, the theories that see reasons as facts can be divided into "realistic" and "irrealistic" ones, and second, that non-psychologistic theories are objectivist and not perspectivist. In the first case, realistic theories support a view of the facts in "some ontologically substantial or robust sense," seeing facts "as the truth-makers for true normative propositions" (Robertson 2009, 13),[56] while "irrealist" theories "deny that such truths are made true by, or obtain in virtue of there

[53] In this sense, Dancy (1993, 32) notes that "what motivates is the matter of fact believed" and that "what motivates is the fact that one believes, which is still a fact."

[54] Alvarez (2016a) points out that there is "disagreement about what facts of any kind are: are they concrete or abstract entities? Is a fact the same as the corresponding true proposition, or is the fact the 'truth-maker' of the proposition? Are there any facts other than empirical facts, e.g. logical, mathematical, moral or aesthetic facts?"

[55] Raz (1999a, b, 18) adds that "a fact is that of which we talk when making a statement by the use of sentences of the form 'it is a fact that…' In this sense facts are not contrasted with values, but include them."

[56] These theories can be distinguished into naturalistic and non-naturalistic ones.

being, normative properties construed as robust or substantial ontological items" (ibid.).[57] Non-psychologistic positions are "'objectivist' in that they presuppose that whether an agent has an (objective) normative reason to act depends solely on the facts and not on the agent's beliefs" (Alvarez 2016a).

4 Epistemic and Practical Reasons

As we have seen, in the sphere of normative reasons it is possible to distinguish reasons in relation to the type of behaviour they may favour. The distinction between epistemic and practical reasons (for action) is based on the type of φ-ing they refer to: the first are those on which basis we "believe something" and/or "make a […] cognitive step within a train of reasoning to a conclusion" (Skorupski 2010, 35),[58] while the latter are those in which our concern is "to perform an action." Epistemic reasons are those that have the concept of truth as a reference point, while the latter may refer to different values, norms, etc. According to the first, rightness or appropriateness is "related to truth," while according to the second, it concerns "the good, broadly conceived, that is, as related to a variety of values" (Alvarez 2010, 3–4).

It follows that "epistemic reasons are governed by one concern: determination whether the belief for which they are reasons is or is not true," while "reasons for a single action may, and typically are, governed by many concerns" (Raz 2009, 41).[59] This can be seen as the most important difference between these two types of reasons: "practical reasons serve many concerns and epistemic ones can serve only one" (ibid., 43).

This difference can be articulated in different ways: as indicated, a first analysis points out that "reasons for believing are not person-relative," while "reasons for acting and wanting are" (Alvarez 2010, 19). From this perspective, this means that, in spite of several possible objections,[60] they do not in principle depend on the agent's perspective. It follows that "if the fact that p is a reason for someone to believe that q, then it is a reason for *anyone* to believe this, no matter what his or her circumstances and goals are" (ibid.). The reasons for the action instead have the "distinctive feature" that "they are the subject of choice:" in addition of referring to different values, this

[57] Robertson (2009, 13) adds that on these positions, "a normative proposition might be true in virtue of satisfying some truth- or knowledge condition—such as those presented by *formal* criteria like universaliability, convergence commitments, the Categorical Imperative, and so on—which, when satisfied, incur no commitment to there being robustly normative properties within the fabric of the world."

[58] Skorupski (2010, 35) brings up as examples such cases in which we "introduce a supposition, make an inference, or exclude a supposition, etc."

[59] For Raz (2009, 37), epistemic reasons are "truth-related"; that is, they "are reasons for believing in a proposition through being facts which are part of a case for (belief in) its truth."

[60] For the objections, see Alvarez (2010), 20–22.

stands in relation both to the "different means of achieving one's goal" and to the "choice between possible but (at the time) incompatible goals" (ibid., 17).[61]

Joseph Raz (2009) has analysed the two types of reason in relation to values, indicating how, in general, reasons for action refer to values (but not only values),[62] while epistemic reasons "are not similarly connected to values, not even to a single value": it is not possible to say that "there is always value in having a true belief" and not even that "it is always a disvalue to have a false belief" (ibid., 43). There follows what can be seen as a specific difference: "the value-independent character of epistemic reasons" (ibid., 44). Unlike reasons for action, these reasons "are not related to values" (Sobel and Wall 2009, 3): they "do not derive from the value of having that belief in the way that reasons for an action derive from the value of that action," and therefore, "reasons for belief are not provided by values in the way that reasons for action are." This is to say that reasons for action can be seen as "presumptively enough," that is enough to justify an action (if there are several reasons, they could be traced to a single reason) or against an action, while "epistemic [reasons] are not necessarily so" (Raz 2009, 43, 44).[63]

In Raz's analysis, epistemic reasons are "adaptive," while those for action are "practical." In the first case, they are reasons that "mark the appropriateness of an attitude in the agent independently of the value of having that attitude, its appropriateness to the way things are," while the latter are "value-related" reasons (ibid., 46). This distinction is matched by that between standard and non-standard reasons: reasons of the first kind (corresponding to epistemic reasons) "are those which we can follow directly, that is have the attitude, or perform the action, for that reason," while reasons of the second kind (corresponding the practical reasons) are "reasons for an action or an attitude […] such that one can conform to them, but not follow them directly" (ibid., 40). This analysis is aimed at highlighting that normativity is not necessarily bound to values: "the value of a thing provides some reasons, but not all" (Raz 2011, 95).[64] The analysis shows (or should show) that reasons should

[61] Raz (2009, 41) underlines that the possibility of choice regards not only single actions that "can serve or disserve a number of intrinsic values," but also action that refers to a single value that "may serve independent concerns." This happens, for example, when "a single act can advance the welfare of several individuals, when the interest of each of them is a reason, an independent reason, to perform it": in this case, we may face the impossibility of their contemporary practicability or of their possible conflict.

[62] Raz (2009, 43) underlines that "the diversity of concerns manifested in practical reasons is not entirely due to the diversity of values. Diverse values do generate diverse concerns, but so do other factors: for example, being a medically qualified caretaker of sheltered accommodation for disabled people I have a reason to help anyone there who needs insulin injections. There are several such people. So I have a reason to help each of them. Each of these reasons represents an independent concern, and they can conflict with each other, even though they all derive from the same value."

[63] Among the other differences that can be added, there is the one noted by Skorupski (2010, 156, xvii, 40, 41), for whom "epistemic reasons are relative to an epistemic field," that is to a "set of facts knowable to the actors" that create "epistemic dependencies." For Skorupski, "epistemic reasons are relative to their field: whether a subset of facts is an epistemic field constitutes a reason—and how good a reason for belief it constitutes—depends on the other facts of the field."

[64] In an even stronger stance, Raz (2011, 95) has argued that "the difference between practical and epistemic reasons is central to the attempt to understand the normativity of reasons. It defeats any

be seen in the different dimensions (standard, non-standard, adaptive, practical) that the distinction between epistemic reasons and reasons for action makes it possible to emphasize.

5 The Modality (and Strength) of Reasons

One of the central aspects of the analysis of reasons for action is that of evaluating the different normative roles they may have in determining the action of the various subjects. As we have seen, this is linked to the fact that reasons for acting are, in principle, pro tanto reasons: this means that we have reasons for a particular action and reasons against the same action.[65] The possibilities of contrast are wide and at a minimum they refer to the following: the different values an action can express, the choice between two types of actions that have different goals, and the choice of the means with which to achieve a given goal. These possibilities give rise to a phenomenology of reasons that highlights the different modalities (and strengths) they can have.

In an analysis comparing different reasons, it is necessary to distinguish between a conflict between first-order reasons and between first- and second-order reasons: in the first case, this is the contrast between the reasons supporting "different and incompatible courses of action," a contrast that "ordinarily we resolve […] by assessing the relative weight or strength of all the relevant reasons and then deciding in favour of that action which has the greatest overall support" (this is therefore the process of "determining what ought to be done on the balance of reasons") (Perry 1989, 973); in the second case, the reference is a reflective process that determines whether "to act on or refrain from acting on a reason" (ibid.) in relation to reasons of different levels that can orient the weighing process in different ways. In this second case, we have reasons for reasons, such as those which can be determined by the presence of mandatory norms (issued by an authority, be it practical or epistemic).

If we accept the distinction between first- and second-order reasons, we will have two types of conflict between reasons: those between first-order reasons and those between second-order (exclusionary) reasons and first-order reasons.[66]

---

attempt to explain normativity as having to do with the influence of value on us. Epistemic reasons have nothing to do with value."

[65] Alvarez (2010, 15) takes the example of bungee-jumping: "If bungee-jumping is a good thing to do (at least for some people) this is presumably because it is fun, thrilling, exhilarating, and so on—that is, it is good for what might be called a hedonic reason. On the other hand, given the risks involved, there seem to be prudential reasons against bungee-jumping."

[66] Raz (1999a, 35) notes that "description of conflicts of reasons and their resolution […] is one of the most intricate e complex areas of practical discourse."

5.1 Conflict and Weighing Between First-Order Reasons

There is a conflict between first-order reasons when "p strictly conflicts with q relative to X and φ if, and only if, R (φ)p,x and R(φ)q,x, i.e. that p is a reason for x to φ and that q is a reason to refrain from φ-ing" (Raz 1999a, 25): basically, when you have reasons in favour of a particular behaviour and reasons against that behaviour.

In order to analyse this type of contrast (between first-order reasons), it is necessary to (a) define the role of reasons in relation to the context; (b) evaluate the types of conflicting reasons and their characteristics; and (c) consider the various options in relation to the choice between possible actions.

(a) Comparison between first-order reasons requires a preliminary choice about the way in which the reasons can be considered in practical reasoning and in relation to the context. In this sphere, following Dancy's suggestions (2004a, 9, 132, 94), it is possible to choose between two possible approaches: holism, which claims "that a feature which has a certain effect when alone can have the opposite effect in a combination," that is "that a feature that normally counts in favour of a (sort of) action may on occasion not count in favour at all," and atomism, which is "the claim that if a feature is a reason in one case, it must be a reason (and on the same side) wherever it occurs." This means that on a holistic approach, the "context can affect the ability of a feature to make a difference in a new case" (ibid., 7).[67]

(b) The reasons that may collide can be different: a first distinction concerns the comparison between epistemic and practical reasons. As noted, epistemic reasons are "standard," while those practical are "non-standard:" that means that "a conflict between a practical reason to believe p and an epistemic reason to believe not p is not a genuine conflict," since "the two kinds of reason do not compete." In these cases, "the epistemic reason will win out" (Sobel and Wall 2009, 3). However, the possible contrast between different epistemic reasons is a contrast between first-order reasons.[68]

First-order reasons can be categorized differently: a general classification that outlines their different role is the one proposed by Dancy (2004a), who in the context of the analysis of "contributory reasons" distinguishes between favouring reasons (*favourers*), enabling reasons (*enablers*), and intensifying reasons (*intensifiers*). A "contributory reason for action" is a pro tanto (or prima facie) reason: it is "a feature whose presence makes something of a case for acting, but in such a way that the overall case for doing that action can be improved or strengthened by the addition of a second feature playing a similar role" and that "is not necessarily destroyed by the presence of a reason on the other side" (ibid., 15). In this context, favouring

[67] In relation to the role of moral principles, Dancy (2004a) associates these two positions with two general approaches to reasoning, that of "generalism" (reasons depend on general principles) and "particularism" (reasons do not depend on general principles). For Dancy, "normally, particularists are holists and generalists are atomists" (ibid., 9). He also distinguishes holism from nonmonotonic reasoning.

[68] For Alvarez (2010, 14), "when the reason to believe something and the reason not to believe it are of equal strength, then believing either may be right."

reasons (favourers) are those that provide a reason to act in a certain way; that is, they can make a specific action "right or appropriate." To these must be added enabling reasons, which are those that refer to conditions that make it possible to act in a certain way, namely those that, so to speak, determine the necessary conditions for realizing the act.[69] To these two reasons, we must adjoin reasons that intensify a reason (intensifier): these are the ones that "strengthen" a favouring reason. The overall picture of reasons envisions "three sorts of role that a relevant consideration can play: a relevant consideration can be a *favourer/disfavourer*, it can be an *enabler/disabler* for another favourer/disfavourer, and it can *intensify/attenuate* the favouring/disfavouring done by something else" (ibid., 42).

A classification developed more directly in relation to the possible conflict between reasons is the one proposed by Raz (1999a). The starting point is the identification of a *complete* reason: this is a reason that is "indispensable in any logical explication of reason," which implies that

> the fact that p is a complete reason to φ for a person x, if, and only if, either (a) necessarily, for any person y who understands both the statement that p and the statement that x φ's, if y believes that p he believes that there is a reason for x to φ, regardless of what other beliefs y has, or (b) R(φ)p,x entails R(φ)p,y which is a complete reason. (Ibid., 24)

This definition is not immediately apparent and can be explained in the light of the problem it wants to explain, that is "the difference between completing the statement of a reason and […] stating a second reason" (ibid., 23). The statement that "wherever φ-ing would increase human happiness one has a reason to φ" is a complete reason if you add the premise that "human happiness is a value": a reason is of this kind if "the fact stated by any set of premises which entail that there is a reason to perform a certain action is a complete reason for performing it" (ibid., 24–25).[70] An *atomic* complete reason can be defined "as a complete reason which would cease being complete if any one of its constituent parts were omitted" (ibid., 25).

A complete reason includes *operative* reasons and *auxiliary* reasons: in the first case, we have reasons that involve an inference, "such that belief in their conclusions entails having a practical critical attitude while no such attitude is required for belief in their premises" (ibid., 33).[71] They are "any reason if, and only if, belief in its

---

[69] Dancy (2004a, b, 30) explains the different roles with this example: "1. I promised to do it. 2. My promise was not given under duress. 3. I am able to do it. 4. There is no greater reason not to do it. 5. So: I do it." Number 1 is a favourer, while 2, 3, and 4 are enablers (general or specific).

[70] Raz (1999a, 24–25) makes this example: "Suppose John says: wherever φ-ing would increase human happiness one has a reason to φ. Let us assume that Jack denies this. How are we to understand Jack's position? Is he guilty of a mistake in logic? Not necessarily. John does not state a complete reason, though it is easy to see which reason he is invoking. It is that human happiness is a value and that under certain conditions φ-ing increases human happiness. This is his complete reason for φ-ing when those conditions obtain." Of course, Jack can continue to deny that human happiness is a value, but he would make a logical mistake if "the reason of his denial is that values do not always constitute reasons, or that sometimes there will be stronger reasons for not φ-ing despite the fact that it contributes to happiness."

[71] Raz (1999a, 34) stresses that "most operative reasons are either values or desires or interests:" the latter can be called "subjective values," while "values are dubbed objective values" (that is

existence entails having the practical critical attitude": this is the case, for example, if you claim that "respect for persons is a value then there reason for everyone to respect persons," or the fact that my having "promised to φ" means that I "have a reason to φ" (ibid.). Auxiliary (identifying) reasons whose function is "to help to identify the act which there is reason to perform" (e.g. to identify whether a loan can be the most appropriate way to help someone in need) and can be countered "with strength-affecting reasons" (that is, what is more beneficial) (ibid., 34–35). They concretize, so to speak, operative reasons. For Raz, complete and operational reasons must necessarily be together: "every complete reason includes an operative reason and that every operative reason is a complete reason for some action or other" (ibid., 33).

Reasons should be compared: the strength of a reason lies in its "power to override." An assessment of the strength of reasons must take account of any *cancelling* conditions (as in the case in which "a friend has released me from a promise"): they eliminate a reason and therefore do not concern their strength (ibid., 27).

In relation to the strength of reasons, it is possible to identify, in addition to prima facie reason, *conclusive* reasons and *absolute* reasons. In the first case, "p is a conclusive reason for x to φ if, and only if, p is a reason for x to φ (which has not been cancelled) and there is no q such that q overrides p," while, in the second, "p is an absolute reason for x to φ if, and only if, there cannot be a fact which would override it; that is to say, for all q it is never the case that when q, q overrides p" (ibid., 27).[72] This leads to the conclusion that "it is always the case that one ought, all things considered, to do whatever one ought to do on the balance of reasons" (ibid., 36).[73]

A particular role in the balancing process can be assigned to moral reasons. This role, however, should not be seen in the manner of the necessary prevalence of moral reasons, that is "that moral reasons for acting always defeat other reasons," for "it would be sufficient that there was no moral reason *against* *φ*-ing (Alvarez 2010, 16)."

---

that "everyone has an operative reason to promote"). On the distinction between subjective versus objective and relative versus neutral, see note 33 above.

[72] Raz (1999a, 28) notes that "not every conclusive reason is absolute. A reason may be conclusive because it overrides all the existing reasons which conflict with it and yet not be absolute because it would be not override a certain possible reason, had it been the case."

[73] In relation to the process of weighing, Parfit (2011, 32–34) distinguishes between *decisive* and *sufficient* reasons: we have a decisive reason "if our reasons to act in some way are stronger than our reasons to act in any of the other possible ways, these reasons are *decisive*, and acting in this way is what we have *most reason* to do. If such reasons are much stronger than any set of conflicting reasons, we can call them *strongly* decisive." We have sufficient reasons if "there is […] nothing that we have decisive reasons to do, or *most* reason to do, because we have *sufficient* reasons, or *enough* reason, to act in any of two or more ways." In these cases "our reasons to do something are sufficient when these reasons are not weaker than, or outweighed by, our reasons to act in any of the other possible ways. We might have sufficient reasons, for example, to eat either a peach or a plum or a pear, to choose either law or medicine as a career, or to give part of our income either to Oxfam or to some other similar aid agency, such as Médecins Sans Frontières."

(c) Often, it seems that a contrast between reasons cannot be solved only on the basis of weighing of reasons, as in the case of multiple options. This can happen in the frequent situations in which "people have a variety of options such that it would accord with reason for them to choose any one of them and it would not be against reason to avoid any of them" (Raz 1999b, 99). This may mean that "in some circumstances reasons are optional" (ibid., 94), that is reasons for which "the fact that there are reasons for a certain response make it an eligible, attractive response, but not one which it is wrong not to adopt." They are situations "where neither believing a proposition nor withholding belief will be irrational" (ibid.). The presence of optional reasons can be explained on the basis of "a special optional type" (ibid., 94) or on the basis of the presence of "incommensurable" reasons. In the first case, the reference is to the distinction between *enticing* and *requiring* reasons: the first would be those that "make an option attractive" (ibid., 100) but do not involve irrationality if they are not followed. This seems to suggest that the choice of an option, because enticing, does not take place on the basis of weighing of reasons. In the second case, the presence of optional reasons would depend on the fact that the reasons are incommensurable and, therefore, on reasons that cannot be assessed on the basis of their weight (lacking a common measure): "Reasons which are incommensurate do not defeat each other" (ibid., 101). These aspects can be analysed in two main ways: on the one hand, as Raz suggests, on the basis of an examination of the different factors of the options involved, by trying to assess the incompatible options on the basis of these factors (and therefore in some way weighing them), while, on the other hand, as Dancy (2004b) seems to suggest, by assessing the role of the conditions of implementation and of context.

5.2 *First-Order Reasons and Second-Order (Exclusionary) Reasons*

As noted, a further level of comparison between reasons is that between first-order and second-order reasons for acting. This means that "determining" actions "on the balance of first-order reasons is not the only mode of practical reasoning" to which we can refer (Perry 1989, 913). This is because of the presence of secondary reasons, that is, as Raz (1999a, 39) states, "any reason to act for a reason or to refrain from acting for a reason." We can become aware of their presence through "the detailed examination of conflicts of reason which forces the recognition that different reasons belong to different levels, which fact affects their impact on conflict situations" (ibid., 35). The example is that of a subject (Ann) "looking for a good way to invest her money." From a friend she receives, late in the evening, a proposal for what might be a good investment, but "she has to decide the same evening for the offer" (within midnight). She is undecided: she knows it might be a good investment, but she needs to evaluate it by comparing with another offer she has received before. "All she requires is a couple of hours" to evaluate the two proposals, but she does not because

she is tired, she had a hard day, and she does not feel able to rationally evaluate the proposal. In short, she rejects her friend's offer, not because she believes that it is not good (compared to the other), but only "because she cannot trust her own judgement at this moment" (ibid., 36). In this way, she "claims to be acting for a reason which is not taken into account" in weighing reasons: what is "special" in this case "is, not that she regards her mental states as a reason for action, but that she regards it as a reason for disregarding other reasons for action" (ibid., 37–38).[74]

This example shows that we can have "a reason for not acting on the balance of reasons" and in particular that we can have "a reason to refrain from acting for a reason" (ibid., 39). These types of reasons are the most important second-order reasons: they are *exclusionary* reasons, that is reasons for which we do not take into account the weighing of first-order reasons (for or against). This shows that there may be a "conflict between a first-order reason and a second-order exclusionary reason" (ibid., 40). What is specific to this type of conflict is that "by a general principle of practical reason […] exclusionary reasons always prevail, when in conflict with first-order reasons" (ibid.).

Exclusionary reasons "may vary in scope; they may exclude all or only some of the reasons which apply to certain practical problems," and they "may also conflict with and be overridden by another second-order reason" (ibid.). The presence of exclusionary reasons shows that "there are two ways in which reasons can be defeated:" by "conflicting" or by "exclusionary" reasons. This possibility raises the problem of how to "distinguish between the two ways in which a reason can be defeated," that is to "have a test" by which to identify the two types of reasons. This problem could create some difficulties,[75] but it is possible to identify cases in which the individuation of exclusionary reasons is totally clear: this is true for "decisions and mandatory norms [that] can only be explained with reference to exclusionary reasons" (ibid., 41).[76]

As noted, in a conflict between exclusionary and first-order reasons, the first "always prevails," though it can "be cancelled by cancelling reasons."[77] We can also mention the fact that "the scope of exclusionary reasons can be affected by […] scope-affecting reasons": since "the scope of an exclusionary reason is the class of reasons it excludes," scope-affecting reasons are those that can strengthen (or narrow the scope of) an exclusionary reason (like the high ranking of an authority) (ibid., 46–47). There can also be conflicts "between second-order reasons" that, like

[74] Raz (1999a, 37–39) offers two more examples, relating to the commands of an authority and to a promise.

[75] Raz (1999a, 45) notes that the conflict between first-order and second-order reasons is characterized by "mixed reactions."

[76] Raz (1999a, 47–48) underlines that there are "two main types of exclusionary reasons": "incapacity-based exclusionary" ones and, in general, "authority-based reasons."

[77] The rule of practical reasoning is, in these cases, that "one ought not to act on the balance of reasons if the reasons tipping the balance are excluded by an undefeated exclusionary reason" (ibid., 40).

first-order reasons, can be considered in the light of the strength of the contrasting reasons.[78]

6 Reasons in (Philosophy of) Law

Reflection on the relationship between legal norms and reasons is largely owed to Joseph Raz's work.[79] This reflection, which has many aspects to it,[80] has largely focused on the question of what kind of reason legal rules are.[81] In the philosophy of law, this question has been tied not only to the identification of the type of reason expressed in legal norms, but also to that of legal normativity (and the relation between law and morals).[82]

In Raz's work (1999a), the identification of what types of reasons are legal norms is primarily developed in relation to mandatory norms[83] and on the basis of "content-independent" considerations.[84] The basic thesis is that these norms and some rules

[78] Raz (1999a, 47) emphasizes that he considers exclusionary reasons as the most important second-order reasons and that he does not discuss "second-order reasons to act for a reason."

[79] Redondo (1999, 98) underlines that, for Raz, "in order to account for the concept of legal norm, one must first have a concept of reason for action."

[80] Redondo (1999, 97) stresses that "the concept of reason for action is thought to be relevant for the study of a broad range of questions. In general, it is considered useful for improving the approach to and the explanation of many controversial issues in the field […]. This is the case, for instance, with respect to the problems of normative authority, the existence or validity of a rule, the way how these affect the reasoning of their addressees, etc."

[81] Another important problem, but mainly related to the theory of law, is that of the relation between reasons and law and in particular that of the role that specific legal reasons (such as those related to rights) may have in the processes of applying the law: in this field, the main issue is that of the priorities, in relation to the content of rights (in the form of a material or axiological hierarchy of such content). This, for example, is the case with Ronald Dworkin (1977) and with his theory of rights as "trumps." As is well known, Dworkin argued that, in the processes of applying the law, individual rights always prevail over the community's general goals: in Dworkin's terms, principles (which express individual rights) always prevail over policies. For the analysis of these aspects, see Bongiovanni and Valentini chapter 5, part III, this volume, on "Balancing, Proportionality and Rights."

[82] This problem, which can be seen as the "classic" one of the philosophy of law, will not be analysed in this contribution.

[83] Raz (1999a, 49) states that he prefers "'mandatory' to the more common 'prescriptive.'" The latter "is often used to characterize a type of meaning or a type of speech act […]. 'Prescriptive' also connotes the presence of someone": these are aspects that do not pertain to rules and principles.

[84] Raz (1999a, 51, 50) states that "one is […] forced to look to content-independent features of rules to distinguish rules from reasons which are not rules." With reference to von Wright (1963, Chap. 5), Raz identifies "four elements in every mandatory norms: the deontic operator; the norm subjects, namely the persons required to behave in a certain way; the norm act, namely the action which is required of them; and the conditions of application, namely the circumstances in which they are required to perform the norm action."

are exclusionary reasons[85]: "the notion of exclusionary reasons is essential to the explanation of mandatory norms, especially in order to understand the ways in which their role in practical reasoning differs from that of ordinary reasons for actions" (ibid., 73–74). More specifically, as we will see, "a mandatory norm," for Raz, "is either an exclusionary reason or, more commonly, both a first-order reason to perform the norm act and an exclusionary reason not to act for certain conflicting reasons" (ibid., 58). This double aspect qualifies norms as *protected* reasons, that is "a special kind of reason which combines a first-order reason with an exclusionary reason" (Redondo 1999, 115).[86]

The demonstration of the direct relation between exclusionary reasons and norms takes place in successive steps: firstly, in reference to rules of thumb and those issued by an authority, Raz shows what justifies the fact that they are exclusionary; secondly, the role of the rules is associated with decisions, and this makes it possible to highlight the role played by acceptance and beliefs; thirdly, in the light of the problem of the distinction between rules and norms, Raz analyses the problem of existence/validity of mandatory norms and their characteristics.

What makes exclusionary reasons the rules of thumb and those issued by authority is their role in practical reasoning: they have a precise function that justifies their use. In the first case, what justifies their use is that they are "labour- and time-saving devices [and] error-minimizing devices" (Raz 1999a, 74). They are therefore justified by the task they carry out as tools in relation to "what ought to be done," to save time and work and reduce risks. They are, then, "reasons for having rules," and as such, they "determine the nature of the rules themselves" (ibid., 59): these are specified in the conditions of application of the exclusionary rule. It is necessary to distinguish between the maxim of experience and rules, as "following a rule entails its acceptance as an exclusionary reason for not acting on conflicting reasons even though they may tip the balance of reasons" (ibid., 61). Regarding norms issued by authority, their role is determined by the "nature of authority," which can be "epistemic" or established to ensure social cooperation. In the first case, the authority is "based on knowledge and experience" that "ought to be followed […] when the advice is based on information or experience which the adviser" owns and which we do not or cannot have. This advice is "justified by the wisdom of the authority" (ibid., 63, 74). The rules based on the requirements of social cooperation are justified by the need to ensure such cooperation and are in this sense necessary: in order to cooperate, there must be exclusionary reasons that enable social actors to act on the basis of the instructions

[85]This consideration starts from the criticism of Hart's (1961) practice theory of norms. For Raz (1999a, 53), "the practice theory suffers from three fatal defects. It does not explain rules which are not practices; it fails to distinguish between social rules and widely accepted reasons; and it deprives rules of their normative character."

[86]For Enoch (2014), this type of reason is a "combination of reasons": "A *protected* reason, as I understand it, is such a combination of reasons: If you have a protected reason to φ then you have both a reason to φ and an exclusionary reason excluding at least some of the reasons against φ-ing." In the same essay, he adds to exclusionary reasons *quasi-exclusionary* reasons, which broaden the range of action of the first (for instance, "they include […] reasons not to deliberate in some ways on some reasons").

of the authority. As a result, conceptually, "norms justified by the need to secure coordination must be regarded as exclusionary reasons" (ibid., 74).

To emphasize a further important aspect of exclusionary reasons, Raz uses the analogy with decisions.[87] They are reasons: "a decision is always, for the agent, a reason for performing the act he has decided to perform and for disregarding further reasons and arguments. It is always both a first-order and an exclusionary reason" (ibid., 66). A decision made by a given situation (such as not carrying acquaintances if my car has mechanical problems) can be generalized and become a rule: in this case, it becomes a general exclusionary reason and therefore no longer linked to the evaluation of the various reasons.[88] This role, which "does not depend on whether [someone] came to follow it one way or the other," is based on the fact that "we […] believe that we are justified in following the rule," that is that the rules are believed to be "valid" reasons for the action (both as first-order reasons and as exclusionary ones) (ibid., 72, 75). Mandatory norms show the same features:

> a person follows a mandatory norm only if he believes that the norm is a valid reason for him to do the norm act when the conditions for application obtain and that it is a valid reason for disregarding conflicting reasons, and if he acts on those beliefs. Having a rule is like having decided in advance what to do. (Ibid., 72–73)

The analogy with decisions leads us to underscore the role of belief in the validity of the rules and so the need to "explain what it means for a mandatory norm to be valid" (ibid., 73).[89]

The analysis of this aspect introduces as reference to the mandatory norms "those who believe in their validity" (ibid., 74), i.e. the participants in the system, and sees the fact of following the rule as a "clue" to their validity; however, the dimension of validity is not linked exclusively to that of belief or acceptance, but remains associated with the justification of exclusionary reasons, that is to their ability to simplify decisions (to be time- and work-saving, etc.) and to the fact that they are produced by authorities.[90] To be valid means both to be able, in given circumstances, to carry out the first task and to come from a "legitimate" authority.[91] In the latter case,

---

[87] For Raz (1999a, 71), "this analogy provides a key to an understanding of the nature of mandatory norms."

[88] Raz (1999a, 73) states that "when the occasion for action arises one does not have to reconsider the matter for one's mind is already made up. The rule is taken not merely as a reason for performing the norm act but also as resolving practical conflicts by excluding conflicting reasons. This is the benefit of having rules."

[89] For Raz (1999a, 73), the reference problem is that "not every rule is a valid reason."

[90] Raz (1999a, 73) seems to merge two levels: the validity of a norm involves "more than […] following a rule" (such that "a norm is valid if, and only if, it ought to be followed") and, at the same time, for a norm to be valid requires that "a person follows a rule only if he believes it to be both a valid first-order and an exclusionary reason," that is a "combination of reasons in the validity of which he believes."

[91] The idea of legitimate authority was developed by Raz in his "service conception" (see Raz 1979; 1986). For a review of the various analyses and criticisms of Raz's authority theory, see Ehrenberg (2011). The problem of authority will not be discussed here: for the analysis of this problem, see Himma and Rodriguez Blanco chapter 8 and 9, part I, this volume, on "Authority" and on "The Authority of Law."

the connection between the origin of the rule and its exclusive character is conceptual because, by definition, its coming from authority determines its exclusivity: "to regard somebody as an authority is to regard some of his utterances as authoritative even if wrong on the balance of reasons" (ibid., 65).[92]

In the attempt to "deflate" (Bix 2011, 412), the debate on the normativity of law, David Enoch proposes a different analysis of the relation between legal rules and reasons. In particular, Enoch analyses what it means that "the law […] gives reasons for action" and reframes the problem of "reason-giving force of the law" (Enoch 2011, 2). This aspect is explored "in the context of a more general theory of reason-giving" (ibid., 3). Enoch identifies three types of reasons for action: *purely epistemic*, *triggering*, and *robust* reason-giving. In the first case, an epistemic reason concretizes when you "indicate to me, or show me, a reason that was there all along, independently of your giving it to me" (ibid., 4), that is when you "call our attention to a reason for action that already applies to us" (Bix 2011, 412).[93] A triggering reason is given when "certain changes in non-normative facts can trigger reasons that already apply to us" (ibid., 413). The example is about the possible reduction in your milk consumption if "your neighbourhood grocer raised the price of milk" (Enoch 2011, 4). Such an increase could be seen as the reason prompting you "to reduce your milk consumption." That is not so for Enoch, who claims that there is no new reason, simply the grocer's manipulation of "non-normative circumstances in such a way" as "to trigger a dormant reason that was there all along, independently of the grocer's actions. Arguably, you have a general reason (roughly) to save money" (ibid.).[94] The third type of reason, robust-giving, "a distinct phenomenon" (ibid.) that comes about when "someone's statements or actions do not simply remind us of existing reasons, or trigger the effect of existing reasons, but create reasons that were not there before:" this is true in particular of "requests and commands" and of "promises or plans" (Bix 2011, 413).

---

[92] The correlation between authority and normativity has been questioned by the authors who support the thesis of the connection between law and morality. Nino (1985), for example, has argued that to derive duties from legal norms, working from a positivistic view, is to lapse into the naturalistic fallacy. As noted by Redondo (1999, 112), "on this basis, Nino concludes that it is reasonable to hold, against the thesis of authors like Joseph Raz, that legal provisions do not express operative reasons for justifying decisions, except when they are identified as moral judgments." In the contemporary philosophy of law of positivist style, Raz's theory of reasons has been almost exclusively discussed in relation to his conception of authority and not in relation to his conception of reasons. Even those, as Essert (2013) who criticizes the idea that law is directly reason-giving, accept the distinction between first-order and second-order (exclusionary) reasons and see the latter as "reasons to deliberate about other reasons in particular ways." In this perspective, "to be normative," law ("legal obligations") needs "to make a difference in the structure or content of our deliberations:" this is possible on the basis of "considerations which make a practical difference in our deliberations without themselves being reasons for action: second-order reasons" (ibid., 27, 30).

[93] Bix (2011, 412–413) explains epistemic reasons through this example: "before I do something rash, you might remind me of my obligation to be a good role model to my child or to my students. This reason was always present, and your reminding me did not in any way change the reasons for action that apply to me, but you effectively helped me to (re-)discover those already-existing reasons."

[94] Enoch (2011, 5) underlines that "examples of this triggering case are all around us."

The decisive step in Enoch's argument is to consider legal reasons not as robust-giving (i.e. as providing new reasons for action) but as triggering reasons.[95] As noted, "he sees no basis for assuming that law always (or "necessarily") gives reasons for action (other than "*legal* reasons for action") (Bix 2011, 214).[96] In essence, Enoch greatly limits the role of law as a reason by reducing it to "non-normative 'triggers' to reasons for action that were always already there" (ibid.). However, he associates this reflection with that of Raz (2006, 1012–1013; 1020) and argues that "in the context of a discussion of authority—plausibly a particular instance of robust reason-giving—Raz […] clearly thinks that the reason-giving involved is an instance of (what I call) triggering reason-giving" (Enoch 2011, 10). In this way, the possibility of giving reasons is shifted to authority and, in some ways, to the reasons why one ought to obey authority.

7 Concluding Remarks

The analysis carried out here has sought to provide a picture of the different types of reasons and, with reference to normative ones, of the source of their capacity to provide reasons for action. Of normative reasons (which can be seen as different ontological entities—facts or mental states), it has been privileged the "weighing conception" (and the vision of the reasons as a pro tanto ones) as it seems the most appropriate in relation to the practical reasoning (and to the legal one). In the philosophy of law, this conception was accompanied by the identification of secondary reasons that should allow the explanation of the normative nature of law and of its authority. In this context, however, the distinction between first-order and secondary reasons was generally accepted and the analysis and research have been mainly focused on the problem of authority of law.

References

Alvarez, M. 2010. *Kinds of reasons: An essay on the philosophy of action*. Oxford: Oxford University Press.

Alvarez, M. 2016a. Reasons for action: Justification, motivation, explanation. In *The Stanford encyclopedia of philosophy,* ed. E. Zalta. https://plato.stanford.edu/archives/win2016/entries/reasons-just-vs-expl.

Alvarez, M. 2016b. *Reasons for action, acting for reasons, and rationality*. https://kclpure.kcl.ac.uk/portal/files/46560247/art_3A10.1007_2Fs11229_015_1005_9.pdf.

[95] Enoch (2011, 8) seems to reduce almost all robust reason-giving to triggering reason-giving: "we must conclude that any case of robust reason-giving is really a case of the triggering of a conditional reason." For a critical examination of Enoch's analysis, see Rodriguez-Blanco (2013).

[96] Enoch (2011, 15) underlines that "we must distinguish, in particular, between the claim that the law gives *legal* reasons for action and the claim that the law gives (genuine, unqualified) reasons for action."

Audi, R. 2010. Reasons for action. In *The Routledge companion to ethics*, ed. J. Skorupski. Abington, Oxon: Routledge.
Bix, B.H. 2011. The nature of law and reasons for action. *Problema. Anuario de Filosofía y Teoría del Derecho*, 5: 399–415. http://www.redalyc.org/pdf/4219/421940003018.pdf.
Blackburn, S. 2010. *The Majesty of reason*. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S0031819109990428.
Bratman, M. 1987. *Intention, plans, and practical reason*. Cambridge: Cambridge University Press.
Bratman, M. 2014. *Shared agency. A planning theory of acting together*. Oxford: Oxford, Oxford University Press.
Broome, J. 2004. Reasons. In *Reason and value: Themes from the moral philosophy of Joseph Raz*, ed. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, 56–90. Oxford: Oxford University Press.
Broome, J. 2013. *Rationality Through Reasoning*. Oxford: Wiley-Blackwell.
Buckareff, A. 2014. Review of Reasons and causes: Causalism and anti-causalism in the philosophy of action, ed. G. D'Oro and C. Sandis. Basingstoke: Palgrave Macmillan, 2013. *Notre Dame Philosophical Reviews*. http://ndpr.nd.edu/news/reasons-and-causes/.
Chang, R. 2004. Can desires provide reasons for action? In *Reason and value: Themes from the moral philosophy of Joseph Raz*, ed. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, 56–90. Oxford: Oxford University Press.
Crisp, R. 2014. Keeping things simple. In *Weighing and reasoning: Themes from the philosophy of John Broome*, ed. I. Hirose, and A. Reisner, 140–154. Oxford: Oxford University Press.
Dancy, J. 1993. *Moral reality*. Oxford: Blackwell.
Dancy, J. 2000. *Practical reality*. Oxford: Clarendon Press.
Dancy, J. 2004a. *Ethics without principles*. Oxford: Oxford University Press.
Dancy, J. 2004b. Enticing reasons. In *Reason and value: Themes from the moral philosophy of Joseph Raz*, ed. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, 1–18. Oxford: Oxford University Press.
Darwall, S. 1992. Internalism and agency. *Philosophical Perspective (Ethics)* 6: 155–174.
Davidson, D. 1963. Actions, reasons, and causes. *The Journal of Philosophy* 60 (23): 685–700.
Dworkin, R. 1977. *Taking rights seriously*. Cambridge, MA: Harvard University Press.
Ehrenberg, K. 2011. Critical reception of Raz's theory of authority. *Philosophy Compass* 6: 777–785.
Enoch, D. 2011. Reason-giving and the law. *Oxford Studies in the Philosophy of Law* 1: 1–38. https://ssrn.com/abstract=2607030.
Enoch, D. 2014. Authority and reason-giving. *Philosophy and Phenomenological Research* 89: 296–332. https://ssrn.com/abstract=2606995.
Essert C. 2013. *Legal obligation and reasons*. https://law.queensu.ca/sites/webpublish.queensu.ca.lawwww/files/files/Faculty&Research/FacultyProfileDocuments/EssertLegalObligationandReasons.pdf.
Finlay, S., M. Schroeder, 2012. Reasons for action: Internal vs. External. *The Stanford encyclopedia of philosophy,* ed. E. Zalta. https://plato.stanford.edu/entries/reasons-internal-external/.
Hart, H.L.A. 1961. *The concept of law*. Oxford: Clarendon.
Hieronymi, P. 2005. The wrong kind of reasons. *The Journal of Philosophy* 9: 437–457.
Hieronymi, P. 2013. The use of reasons in thought (and the use of earmarks in arguments). *Ethics* 1: 114–127.
Horty, J.F. 2007. Reasons as defaults. *Philosophers' Imprint* 3: 1–26. http://www.umiacs.umd.edu/~horty/articles/007003.pdf.
Mele, A. 2003. *Motivation and agency*. Oxford: Oxford University Press.
Nagel, T. 1970. *The Possibility of altruism*. Princeton: Princeton University Press.
Nino, C.S. 1985. *La validez del derecho*. Buenos Aires: Astrea.
O'Connor, T. 2010. Reasons and causes. In *A companion to the philosophy of action*, ed. T. O'Connor, and C. Sandis, 129–138. Oxford: Wiley-Blackwell.
Parfit, D. 1984. *Reasons and persons*. Oxford: Clarendon Press.
Parfit, D. 2011. *On what matters*, vol. 1. Oxford: Oxford University Press.

Perry, S.R. 1989. Second-order reasons, uncertainty and legal theory. *Faculty Scholarship. Paper* 1354. http://scholarship.law.upenn.edu/faculty_scholarship/1354.
Raz, J. 1979. *The authority of law: Essays on law and morality*. Oxford: Oxford University Press.
Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon.
Raz, J. 1999a. *Practical reasoning and norms*. Oxford: Oxford University Press (1st ed., 1975).
Raz, J. 1999b. *Engaging reason: On the theory of value and action*. Oxford: Oxford University Press.
Raz, J. 2006. The problem of authority: Revisiting the service conception. *Minnesota Law Review* 9: 1003–1044. https://ssrn.com/abstract=999849.
Raz, J. 2009. Reasons: Practical and adaptive. In *Reasons for action*, ed. D. Sobel, and S. Wall, 37–57. Cambridge: Cambridge University Press.
Raz, J. 2011. Reason, rationality, and normativity. In Id., *From normativity to responsibility*. Oxford: Oxford University Press.
Redondo, M.C. 1999. *Reasons for action and the law*. Dordrecht: Springer.
Ridge, M. 2011. Reasons for action: Agent-neutral vs. Agent-relative. *The Stanford encyclopedia of philosophy,* ed. E. Zalta. https://plato.stanford.edu/archives/win2011/entries/reasons-agent.
Ryle, G. 1949. *The concept of mind*. London: Hutchinson.
Robertson, S. 2009. Introduction: Normativity, reasons, rationality. In *Spheres of reason: New essays in the philosophy of normativity*, ed. S. Robertson, 1–28. Oxford: Oxford University Press.
Rodriguez-Blanco, V. 2013. Reasons in action v Triggering-reasons: A reply to Enoch on reason-giving and legal normativity. *Problema: Anuario de Filosofía y Teoría del Derecho* 7: 3–25. http://www.redalyc.org/articulo.oa?id=421940005001.
Scanlon, T.M. 1998. *What we owe to each other*. Cambridge, MA: Belknap Press of Harvard University Press.
Scanlon, T.M. 2014. *Being realistic about reasons*. Oxford: Oxford University Press.
Schroeder, M. 2007. *Slaves of the passions*. Oxford: Oxford University Press.
Searle, J. 2001. *Rationality in action*. Cambridge, MA: The MIT Press.
Smith, M. 1987. The humean theory of motivation. *Mind* 96: 36–61.
Skorupski, J. 2010. *The domain of reasons*. Oxford: Oxford University Press.
Sobel, D., and S. Wall. 2009. Introduction. In *Reasons for action*, ed. D. Sobel, and S. Wall, 1–12. Cambridge: Cambridge University Press.
von Wright, G.H. 1963. *Norm and action: A logical enquiry*. London: Routledge and Kegan Paul.
White, A. (ed.). 1968. *The philosophy of action*. Oxford: Oxford University Press.

Reasons in Moral Philosophy

Carla Bagnoli

While the concept of reason is pervasive in our ordinary practices, there is a large and divisive disagreement about the role of reasons in moral philosophy. Such disagreement depends on three related issues, which concern the definition of "moral reasons," their sources, and functions.

1 What Is a Moral Reason?

As a working hypothesis, let us establish that a reason is a consideration that counts in favor of something (Scanlon 1998). Typically, there are some facts that count toward believing or acting in a given way. The fact that Laura is an adult Italian citizen counts in favor of believing that she holds voting rights in Italy, and this counts in favor of treating her as entitled to vote in the upcoming elections. What is a *moral* reason? There are two ways to approach this question. On the *material definition*, a reason is moral insofar as it represents some moral facts or moral properties. That is, reasons are moral because their contents are distinctive and peculiar; i.e., they are specifically moral contents. For instance, considerations about harming people or enhancing persons, respecting or undermining their autonomy are typically considered to be moral reasons. How to account for the specific content, import, and strength of such claims is a question for normative ethics. For instance, Kantian theories prohibit harm on the basis of the principle of respect for the dignity of humanity. Utilitarian theories prohibit harm insofar as it does not maximize the utility of all sentient beings (Mill

C. Bagnoli (✉)
Dipartimento di Studi Linguistici e Culturali, Università di Modena e Reggio Emilia, Modena, Italy
e-mail: carla.bagnoli@unimore.it

C. Bagnoli
University of Oslo, Oslo, Norway

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_2

1998). According to virtue ethics, malice is a vice that undermines human flourishing. Such theoretical accounts provide different justifications for their claims, but they often converge on the kind of considerations regarded as morally relevant.

On the *formal definition*, instead, moral reasons are not identified by specific moral contents, but by their internal structure (Frankena 1958). For instance, according to Kantian theories, a consideration counts as a moral reason if it can be the agreed by ideal agents engaged in the activity of co-legislation. This definition does not separate moral reasons from other sorts of reasons, such as epistemic reasons for believing a certain proposition or aesthetic reasons for valuing a certain object. In fact, the underlying claim is that there is no sharp line dividing moral and non-moral reasons. Universality is the formal property of all kinds of reasons, which means that to reason at all we have to be guided by some principles. This is not to deny that there are some moral contents that qualify as moral reasons. The point is that such contents count as moral reasons because of some structural feature such as their function and constitutive aim, rather than in virtue of the fact that they represent a portion of reality. The appeal to the formal structure of moral reasons evokes the metaphysical contrast between form and matter, but it may be interpreted in a way that does not commit to any metaphysics. On this reading, the appeal to the structural arrangement of reason indicates that some sort of consideration counts as a moral reason in virtue of its rational justification. Since rational justification is justification by principles, to count as a reason, any consideration must be principled (i.e., universalizable).

The formal and material approaches to the definition of reasons may converge on the scope of moral reasons, but they importantly differ as to how to account for what makes a consideration a moral reason. For instance, virtue ethics and Kantian ethics agree that there is a moral reason to tell the truth, but the former holds that this is because truth-telling represents a virtue, and the latter holds that this is because lying cannot serve as a viable principle of a universal co-legislation. More fine-grained distinctions emerge when we take into account the more specific functions of reasons.

2 Explanatory and Normative Reasons

Typically, we invoke moral reasons to explain or justify our actions, attitudes, or beliefs. Explaining and justifying are two basic functions of moral reasons, and correspondingly, we may draw the following distinction. *Explanatory* reasons are reasons that make one's attitude or action intelligible to ourselves and to others. *Normative* reasons are considerations that and guide the agent in deciding what to do. For present purposes, we refer to actions in a broad sense, which is inclusive of all cognitive and affective rational activities in which we engage. Beliefs, feelings, emotions, and attitudes belong to this category, even though their constitutive aims and criteria of success differ from the ones pertaining to the performance of an action. Feelings and beliefs are neither decided nor deliberated in the same ways some actions are, but they can be reflectively endorsed on the basis of reasons.

The distinction between normative and explanatory reasons is not meant to be mutually exclusive. In fact, in order to explain action, we often cite normative reasons for action. Explanatory reasons thus often serve as rationalizations of action (Anscombe 1957). In this case, normative reasons are retrospective and explain past actions rather than guiding prospective actions. According to some, the two sorts of reasons are closely, if not conceptually, connected.

When we attempt to explain an action, we attribute normative reasons to the agent. Typically, observers deploy explanatory reasons and refer to considerations that presumably guided the agent in deciding what to do. For instance, Claire thinks that Fabien refuses to raise a family because he is afraid of losing personal autonomy. The observer's explanation of the action succeeds if it makes the action intelligible on the basis of the agent's reasons for performing it. Claire's consideration makes sense of Fabien's actions.

Of course, Fabien may disagree that his decision to raise a family is based on fear. The observers' explanations do not necessarily coincide with the agent's first-person explanation of the same act. It is a platitude that observers and agents often disagree about the explanatory reasons of the agent's action, but this disagreement is philosophically interesting. On the one hand, agents seem to have a special kind of authority about their own actions insofar as they are their authors. On the other hand, agents do not always know more than their observers about the mechanics of their own action. This is a *problem of opacity* of explanatory reasons. Despite Fabien's protest to the contrary, Claire may be right that fear is the real cause of his decision because Fabien is in a state of denial or is self-deceptive. Claire's attribution is correct, and Fabien is not. To capture these disagreements and investigate their philosophical implications, it is useful to distinguish between *attributive* and *operative* reasons. In this case, Fabien's operative reason is fear.

3 The Issue of Agential Authority

The phenomenon of opacity tells us that observers may be better positioned to attribute correctly causes to agents. In the previous example, Claire appreciates and identifies Fabien's operative reasons better than Fabien himself. However, it is questionable whether observers have the authority to explain actions by correctly identifying their motivational causes. While Claire may correctly classify or understand Fabien's decision to raise a family as caused by fear, it is still Fabien's own exclusive business whether to raise a family or not. This is because normative reasons are importantly related to *reasons for action*. They are important reasons for doing (or omitting) something or for undertaking (or disavowing) some attitudes or other. Fabien's reasons not to raise a family are indeed in a distinctive sense his own reasons. Agents have a special claim on their actions: This is called *first-person or agential authority* (Anscombe 1957; Chap. 4). When we raise questions about what to do, we are engaging action as agents, rather than as bystanders. We are asking for considerations that we endorse as reasons for action. Such endorsement is open

exclusively to the prospective agents of the action, and in this specific sense, agents have a special claim on action. That is, agents exercise a special authority on their actions insofar as actions are theirs. The problem is to explain why and how so.

There are competing views of agential authority. According to Kantian constructivism, agential authority implies autonomy, the capacity for self-government, and normative guidance by principles. On an alternative view, autonomy is afforded by reflective endorsement, which is neither principled nor deliberate. Rather, it is a granted by discrete acts of wholehearted identification, which determine the will as well as the bounds of the person (Frankfurt 1988, 1999).

There are also different views about where to situate the philosophical problem of agential authority. According to G.E.M. Anscombe, the problem is both epistemic and practical. The first-person perspective is the perspective of practical knowledge, as opposed to the speculative knowledge of action as an outward performance (Anscombe 1957). By contrast, Christine M. Korsgaards insists that the role of normative reasons is better understood as deliberative, rather than epistemic (Korsgaard 2008, 310–317). It is at the time of deliberation that the question arises whether reasons are efficacious. Normative reasons exert their authority directly, through action itself (Korsgaard 2008, 317). Instead of asking how to put normative truths in practice, it is better to focus on the mechanics of authoring and authorizing action, which pertains to the agent's own perspective.

The issue of agential authority is rooted in the subjective perspective of the agent. Therefore, it raises important issues about the objectivity of reasons for action, their impersonality and impartiality.

4 Subjective and Objective Reasons

Subjective normative reasons are those considerations that the agent takes as relevant because of her partial understanding of the situation, driving interests, and limited information. For instance, Marc justifies his policy of raw-eating on the basis of two kinds of considerations: He does not like cooking and believes that this policy is more ecological in that it has no impact on the planet. These are Marc's subjective reasons for eating raw food. In case Marc's belief that raw food is more ecological is correct, this consideration is also an objective reason to eat raw. However, if this information turns out to be incorrect and actually Marc is wrong in believing that the policy of eating raw food has no impact on the planet, his subjective reasons may still hold. That is because, subjectively, he may still have subjective reasons to eat raw food given what he knows. He may simply be misinformed.

The distinction between subjective and objective reasons comes in degrees, and the two classes of reasons are not mutually exclusive; indeed, they may overlap in most cases. More complex are those cases where subjective and objective reasons prescribe incompatible courses of action. If Marc has a strong preference for raw food, but this kind of food is not healthy, he is criticizable, although he may not be

blameworthy for his false belief. A further question is, whether there is a reason to question his position and correct him and if so on which authority.

It might seem that it is a requirement of rationality that subjective reasons should be abandoned or revised when they are shown to be wrong. However, there is a deeper and more general issue at stake concerning the relation between objective and subjective reasons, especially when values are involved. Subjective reasons are in some important sense "reasons." They are not merely illusory or defective. Rather, they are considerations that spring from the agent's own interests and understanding of the situation. For instance, even though Fabien misunderstands the role of fear in his own situation, his subjective reasons may coincide with the operative reason not to raise a family. His account of the situation is perspectival and yet pragmatically fit.

Because rational justification is driven by the quest for objectivity, it is often ignored that partiality plays a large role in the rational assessment of the situation (Nagel 1986). As it emerges in the case of epistemic subjective reasons, however, speaking from a specific perspective is not necessarily an epistemic defect, but an evaluative component of some kinds of reasons (ibid. Elgin 2017). In fact, some core moral values are for their own nature partial and perspectival and such that they do not command universal endorsement. Perhaps, the most paradigmatic case is love. Reasons for love are distinctively special and partial (Frankfurt 2004). They are not meant to elicit universal endorsement. Reasons for love may be said to exhibit a very peculiar singularity, and a common proof for this claim is that lovers are invariably at loss to explain their love in principled terms. Giorgio may be able to cite some characteristic features of Budapest's baths as reasons for loving Hungary rather than Tuscany, but they would hardly be considerations for convincing anybody else that they should love Hungary rather than any other place exhibiting the same relevant features. Likewise, special commitments and personal bonds generate reasons that are not universal in scope and authority. Reasons that derive from personal commitments, political ideals, friendships, and loving relations seem insulated from the requirement of universal authority that applies in the case of objective reasons. Perhaps more importantly, to adequately account for protecting moral values such as love and friendship, it seems that, the independence of subjective reasons should be preserved.

5 Personal and Impersonal Reasons: Integrity and Authenticity

This is where the distinction between subjective and objective reasons intersects another important distinction between *personal* and *impersonal* reasons. Some moral reasons spring from special relations and are rooted in special concerns. Moral obligations we have to our families, friends, and fellow citizens are of this kind, and their compellingness and import do not seem to compare to obligations we have

toward strangers. For some, these *special* obligations represent the core of morality; for others, morality in the narrow sense coincides instead with the obligations we have toward any other persons, groups or peoples, regardless of special bonds. While moral obligations are often born as burdensome constraints on our action, requests that undercut our projects and concerns, or external impositions that undermine our personality, special obligations are perceived as the very stuff of moral life and appreciated as crucial modes of expressing our integrity and authenticity. Moral decisions and personal preferences are the axes of morality. Personal decisions show that the action is authentically the agent's own action, rather than something imposed from an alien source of authority. This is an important aspect of autonomy, which distinguishes actions that the agent decided to perform because they are expressive of her character and integrity, from those in which she plays only or mostly a causal role, as it happens when the agent is coerced, threatened, or acts unwillingly to avert a greater evil.

To vindicate this aspect, it seems that one should resist the view that objectivity requires the moral agent to overcome the partialities of the personal stance as immoral biases. Furthermore, the presumption that the subjective stance on moral value is a defect to be corrected by broadening its scope and reach an objective standpoint is problematic. Some philosophers argue that the subjective and personal nature of these reasons matter more than their objective vindication (Williams 1981c, Wolf 1997). When moral reasons are completely alien to the agent's own deliberative set or undermine the agent's integrity by undercutting all her subjective reasons, they can hardly exert rational authority over them. This approach raises the question of the place of moral reasons in our life. Critics argue that to attribute overridingness to moral reasons impoverishes character and undermines the richness of personal relations, producing a dull and single-minded agent, or so the objection goes. In response to this objection, philosophers have insisted on the continuity between moral and practical reason (Annas 1993, Engstrom 2009).

6 Drawing the Boundaries of the Moral Domain

How to draw the boundaries of the moral domain is a very controversial philosophical question which has important implications. While we often refer to "common morality," it is arguable that this expression must be understood indexically, that is, as referred to the specific morality held at a specific time and place, by some society. This *descriptive use* of the term "morality" is prevalent in anthropology and comparative and evolutionary psychology (De Waal 1996, Sinnott-Armstrong 2008). The implication of this definition is that there is no universal and universally authoritative body of moral cognitions or moral norms. By contrast, on some *normative use* of the term "morality," it refers to a body of moral cognitions or norms that are available to and binding for all relevant agents. That is, ideally or under specified conditions, all relevant agents are guided by such moral norms. According to the normative understanding, moral norms are typically considered impartial,

overriding non-moral considerations and universally authoritative and binding for all relevant agents (Kant 1967, Darwall 1983, O'Neill 1989). Moral theories differ as to the account of the source of authority and the contents of such moral norms. They also importantly differ as to the inclusiveness of the class of relevant agents. For some, moral norms apply to all rational beings. For others, moral norms apply only to human rational agents because of their peculiar and distinctive epistemic and practical limitations, such as fallibility, frailty, and mutual vulnerability. Supporters of morality hold that it is a cooperative enterprise, which generally favors human flourishing and represents the rationally best way to deal with problems such as distributing scarce resources (Baier 1958, Rawls 1971, Frankena 1973, Gauthier 1986). Debunkers hold that morality is an instrument in the service of some groups, which manipulate their partners in order to promote their own specific interests, or else it is a system of blame provided by natural selection. In any case, moral judgments do not have any special authority, but they are the expression of particular perspectives, visions, or projections In particular, moral judgments lack both ontological support and rational authority, so that their apparent inescapability or necessity does not sustain close investigation (Mackie 1977, Joyce 2001).

7 Moral Reasons and Moral Reasoning

Moral theories not only differ in their account of the contents of moral reasons, but also in their account of how moral reasons are produced or recognized. On the *recognitional view*, moral reasoning aims at recognizing what there is moral reason to do, to believe or to feel. It starts with some moral premises and ends with conclusions about what to do, to feel, or to believe. On the *constructivist* view, instead, reasoning builds up reasons according to some procedure, against the background of a conception of moral agency and relevant facts of the matter. A further important question is whether and how moral reasoning relates to practical reasoning. According to Aristotelian views, moral reasoning is part and parcel of a general account of practical reasoning, which amounts to the specification of ends. The wise is capable of recognizing the good ends because of their adequate upbringing, but the structure of their reasoning is not too different from the vicious persons. The wise and the vicious have different ends, but they determine their ends in a similar manner. However, the wise is capable of harmonious integrity because their ends fully cohere, while the vicious are always at risk of disintegration because their ends cannot integrate and push apart (Engstrom 2009). Kantians offer a unified account of practical reasoning, whose universal structure mirrors theoretical reasoning. However, there is an important distinction between *moral* and *prudential reasoning*. Prudential reasoning is hypothetical because it is conditional on some particular ends that the agent actually holds; hence, it does not produce unconditional obligations that apply to all rational beings. By contrast, moral reasoning does not depend on any specific end, but it is modeled by universal co-legislation. The difference between these two kinds of reasoning concerns their authority. While the conclusions of moral reasoning

hold universally and unconditionally for all relevant agents who are represented as co-legislators, the conclusions of prudential reasoning hold for the narrow group of agents who share in their specific ends or interests (Korsgaard 1997). Furthermore, the authority of prudential or instrumental reasoning depends on the authority of moral reasoning. That is to say that instrumental reasoning carries normative significance only against the background of a general account of practical reasoning. Not all agree that moral reasons are unconditional and necessary, as Kantians do. Humeans hold that all moral reasons are hypothetical, and not structurally different than other sorts of practical reasons, such as etiquette (Foot 1978, McDowell 1978).

8 Moral Reasons in Conflict

Moral reasons are generally thought to exhibit a distinctive kind of gravity and importance. This is often explained with the philosophical claim that they represent an eminent domain of objects, e.g., values of a particularly dignified sort. Moral reasons are both unconditionally authoritative and rationally overriding. This view is widely challenged, as the fortune of Williams' argument about internal reasons shows. The issue of the normative force of moral reasons importantly relates to the phenomenon of moral conflict, which ramifies through personal and interpersonal dimensions.

In the intra-personal case, moral reasons may clash with non-moral or prudential reasons. As the Kantian case of the prudent merchant shows, moral reasons need not be always in contrast to prudential reasons: Sometimes they converge on recommending the same line of action. It might be prudent to price one's merchandize fairly in order to keep one's clients. In other cases, however, it might be more profitable to exploit, and these are cases of moral conflict. According to Kantian theories, moral reasons are rationally overriding in deliberation in that they always trump non-moral preferences, interests, and desires (Kant 1967, Korsgaard 1996). However, this is not to say that interests, desires, and preferences do not provide reasons for action. On the contrary, the claim is that they generally do, and when they clash with moral reasons, they fail to provide definitive reasons for action. Other moral theories admit of alternative ways to treat the relation among moral and non-moral reasons for action, which include overridingness, weakening, outweighing, annulling, defeating, and neutralizing (Nozick 1968, 30–35). When defeated or undermined, moral reasons leave a remainder which generates further moral reasons for compromise, reparation, and compensation. Furthermore, for some moral theories, moral learning through reasoning and experience allows for a variety of ways in which moral and non-moral reasons align and can be made cohere and integrate into time. Moral education is thus an important aspect of a theory of moral reasons in that it broadens and deepens the ways of coping with moral conflicts.

This diachronic dimension of moral reasoning is particularly relevant in the account of interpersonal moral conflicts, where the parties disagree not only because their interests clash but also because their values do. Conflicts of this kind are

pervasive in a pluralistic society, and it is part of democratic agenda to understand how to legitimately address such conflict. If pluralism is a value to be protected and fostered, some conflicts and disagreements ought to be preserved (Williams 1981a). Yet it must be possible to find ways of accommodating such conflicts and disagreements without undermining the civic structure of the society (Rawls 1993, Nozick 1968). In such cases, norms of basic rationality help citizen to deal with conflict of moral reasons, without entering the underlying dispute about values. This view seems to imply that all rational agents share the same basic norms of rationality, which allows that to share a conception of "public reason" (Rawls 1993, Rawls 1999). However, it is arguable that norms of rationality do not guide us independently of commitment to any specific values. In fact, some hold that even norms of rationality deeply depend on more fundamental special identities, and therefore, they are conditional requirements that depend on sharing such special identities (MacIntyre 1988). To fruitfully address this crucial problem, it is helpful to distinguish claims of structural rationality, which depend on how we form reasons, and claims of substantive rationality, which vary according to specific normative theories (Scanlon 2007). This distinction allows us to identify different levels and dimensions of rational disagreements and investigate distinct modes of resolution.

9 Moral Reasons and Coordination

On a prominent view, morality is basically a cooperative enterprise aiming at coordination (Kant 1967, Hobbes 1994, Frankena 1973). This is the point of convergence between Hobbesian and Kantian traditions, insofar as they attempt to found moral obligations as rational requirements. However, Hobbesians take reason to be basically self-interested while Kantians hold that by undertaking reasoning one is also already committed to morality, in the minimal sense of a fundamental disposition to reason with others. Critics have argued that the Kantian solution is question-begging, since it assumes some basic moral commitment. But it is also, and more radically, open to debate that morality works as a coordination device, since moral differences and disagreements are pervasive and divisive. This latter worry can be addressed by distinguishing between some basic moral concern or moral dispositions (such as the requirement to reason with others) and the adoption of a full-fledged substantive morality or moral code. The claim that morality is a coordination device is compatible with the existence of different moral codes, traditions, and also with cultural change over time and institutional transition. Moreover, the emergence of moral codes and practices is amenable to evolutionary explanations (Gibbard 1990, Hauser 2006, Street 2006, Copp 2008, Fitzpatrick 2014). The basic point is that we exchange moral reasons with others in order to solve some coordination problems. To this effect, moral reasons importantly contribute to represent the problem at hand. These reasons concern not only actions to undertake, but also emotions to express and beliefs to endorse.

The importance of moral reasons as cooperative schemes in coordination problems may be taken as an aspect of a more general feature of reasoning. On some prominent views, reasoning is dialogical (Kant 1967, Habermas 1984). If we take moral reasoning to be characterized by universalization, then its structure is always *collective* as it is fundamentally reasoning among different personae. How to design the personae involved in collective moral reasoning is a philosophical question, which carries divisive theoretical and practical disagreements. The dialogical account of moral reasoning can be associated with different models of moral agents, with radically different results. This view importantly implies that there is no natural and trivial answer to the issue of the boundaries and the category of relevant agents involved; hence, the scope of applicability of moral obligations is not naturally defined.

10 Moral Reasons and Compliance

A significant problem related to the category of moral agency to which moral reasons apply concerns the issue of compliance. If moral obligations are rational requirements, is compliance with morality granted by rationality alone? The question is complex, and it ranges over two partially independent debates. The first debate concerns the source of authority of moral reasons and revolves around the distinction reason/sensibility; the second debate concerns motivational force of moral reasons and revolves around the internal/external reasons distinction. As for the debate concerning source of authority of moral reasons, we might distinguish two camps: rationalists and sentimentalists. On the rationalist view, moral reasons apply to all beings endowed with rationality and are universally binding and authoritative. Rational agents take moral reasons to be intrinsically authoritative and compelling. That is, they are guided by what is morally dutiful or virtuous independently of what is merely desirable or prudent. This is not to say that all rational agents always comply with morality, as they might be overpowered by external forces or interfered with. The point is that if all goes well, moral reasons are capable of driving rational action without further motivational aid. According to sentimentalists, instead, reasons are motivationally inert beliefs, which cannot compel action independently of the presence of desires. Alternatively, sentimentalists explain moral motivation in terms of moral sentiments, attitudes, dispositions, and desires that motivate the agent to act according to duty. This disagreement about the motivational import of moral reasons is thus rooted in a difference about the powers and scope of reason (Korsgaard 1996, Smith 2013). However, the question is not solved simply choosing between these two accounts of practical reason. To take morality as a requirement of rationality raises questions about how to make requirements compelling. On the externalist view, moral requirements are compelling insofar as they are combined with a desire to comply (or some other motivational force). This view makes the authority of moral reasons conditional on some external sanctions, such as blame and reprobation or some internal sanctions such as sense of guilt, remorse, or shame. On the internalist view, instead, moral requirements are motivating insofar as they are themselves normative (Nagel

1970). Partly, the question relates to whether reasons—understood as beliefs or principles—can generate desires or motivate action independently of desire (Parfit 1997, 2006). According to Williams (1981b, 101–113), all reasons are internal in that they are necessarily linked to the agent's actual deliberative set. Williams' argument that genuine reasons for action must be linked to what we care about challenges a too narrow and moralistic conception of practical reasoning (Williams 1981b, Frankfurt 2006). The intended effect of the argument is that there are no reasons that apply independently of the agent's special plans, motivations, and projects. However, there is always a question about how to integrate such projects and plans and coordinate with other relevant agents who are equally entitled to carry their own projects and plans. To adequately address such problems, it seems crucial to acknowledge the moral and political varieties of ends at stake in practical reasoning and refocus the debate on the resources for reasoning with others.

References

Annas, J. 1993. *The morality of happiness*. Oxford: Oxford University Press.

Anscombe, G.E.M. 1957. *Intention*. Oxford: Basil Blackwell.

Baier, K. 1958. *The moral point of view*. Ithaca, NY: Cornell University Press.

Copp, D. 2008. Darwinian skepticism about moral realism. *Philosophical Issues* 18: 186–206.

Darwall, S. 1983. *Impartial reason*. Ithaca, NY: Cornell University Press.

De Waal, F. 1996. *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, Mass.: Harvard University Press.

Engstrom, S. 2009. *The form of practical knowledge*. Cambridge, MA: Harvard University Press.

Fitzpatrick, W. 2014. Morality and evolutionary biology. *Stanford Encyclopedia of Philosophy*. Available online at: https://plato.stanford.edu/entries/moralitybiology/.

Foot, P. 1978. Morality as a system of hypothetical imperatives. In *Virtues and vices*, ed. P. Foot. Berkeley, CA: University of California Press.

Frankena, W.K. 1958. Obligation and motivation in recent moral philosophy. In *Essays in moral philosophy,* ed. A.I. Melden, 40–81. Seattle, WA: University of Washington Press.

Frankena, W.K. 1973. *Ethics*. Englewood Cliffs, NJ: Prentice-Hall.

Frankfurt, H. 1988. *The importance of what we care about. philosophical essays*. Cambridge: Cambridge University Press.

Frankfurt, H. 1999. *Necessity, volition, and love*. Cambridge: Cambridge University Press.

Frankfurt, H. 2004. *The reasons of love*. Princeton, NJ: Princeton University Press.

Frankfurt, H. 2006. *Taking ourselves seriously and getting it right*. Stanford, CA: Stanford University Press.

Gauthier, D. 1986. *Morals by agreement*. Oxford: Oxford University Press.

Gibbard, A. 1990. *Wise choices, apt feelings*. Oxford: Clarendon Press.

Habermas, J. 1984. *Theory of communicative action,* trans. T. McCarthy. Boston, MA: Beacon Press. (1st ed., 1981).

Hauser, M. 2006. *Moral minds: How nature designed our universal sense of right and wrong*. New York, NY: Harper Collins.

Hobbes, T. 1994. *Leviathan,* ed. Edwin Curly. Indianapolis, IN: Hackett Publishing Company (1st ed., 1660).

Joyce, R. 2001. *The myth of morality*. Cambridge: Cambridge University Press.

Kant, I. 1967. *Groundwork of the metaphysics of morals.* New York, NY: Barnes & Noble. (1st ed., 1785).

Korsgaard, C.M. 1996. Skepticism about practical reason. In *Creating the Kingdom of Ends.* Cambridge: Cambridge University Press.
Korsgaard, C.M. 1997. The normativity of instrumental reason. In *Ethics and practical reason*, ed. Garrett Cullity, and Berys Gaut, 215–254. Oxford: Oxford University Press.
Korsgaard, C.M. 2008. *The constitution of agency*. Oxford: Oxford University Press.
MacIntyre, A. 1988. *Which justice? Whose rationality*. Notre Dame, Ind.: Notre Dame University Press.
Mackie, J.L. 1977. *Ethics: Inventing right and wrong*. London: Penguin.
McDowell, J. 1978. Are moral requirements hypothetical imperatives? *Proceedings of the Aristotelian Society Supplementary* 52: 13–29.
Mill, J.S. 1998. *Utilitarianism.* ed. Roger Crisp. New York: Oxford University Press. (1st ed., 1863).
Nagel, T. 1970. *The possibility of altruism*. Oxford: Clarendon Press.
Nagel, T. 1986. *The view from nowhere*. Oxford: Oxford University Press.
Nozick, R. 1968. Moral complications and moral structure. *Natural Law Forum* 13 (1): 1–50.
O'Neill, O. 1989. *Constructions of reason,* 206–218. Cambridge: Cambridge University Press.
Parfit, D. 1997. Reasons and motivation. *Proceedings of the Aristotelian Society Supplementary* 71: 99–130.
Parfit, D. 2006. Normativity. In *Oxford studies in metaethics,* ed. Russ Shafer-Landau, vol. 1, 325–380. Oxford: Clarendon Press.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Rawls, J. 1993. *Political liberalism*. New York, NY: Columbia University Press.
Rawls, J. 1999. The idea of an overlapping consensus. In John Rawls, *Collected Papers.* ed. S. Freeman, 421–448. Cambridge, MA: Harvard University Press. (1st ed., 1987).
Scanlon, T.M. 1998. *What we owe to each other*. Cambridge, MA: Harvard University Press.
Scanlon, T.M. 2007. Structural irrationality. In *Common minds: Themes from the philosophy of Philip Pettit*, ed. G. Brennan, R. Goodin, F. Jackson, and M. Smith, 84–103. Oxford: Oxford University Press.
Sinnott-Armstrong, W. (ed.). 2008. The evolution of morality: Adaptations and innateness. In *Moral psychology* 1. Cambridge, MA: MIT Press.
Smith, M. 2013. A constitutivist theory of reasons: Its promise and parts. *Law, Ethics, and Philosophy* 1, pp. 9–30.
Williams, B. 1981a. Conflicts of values. In *71-82*, ed. Moral Luck. Cambridge: Cambridge University Press.
Williams, B. 1981b. Internal and external reasons. *Moral luck*, 101–113. Cambridge: Cambridge University Press.
Williams, B. 1981c. Moral Luck. *Moral luck*, 29–37. Cambridge: Cambridge University Press.
Wolf, S. 1997. Moral saints. In *Virtue ethics,* ed. R. Crisp and M. Slote, 79–98. Oxford: Oxford University Press. (1st ed., 1986).

Legal Reasoning and Argumentation

Douglas Walton

Wigmore (1931) thought that there was a science of proof underlying legal reasoning. He thought this science of proof was inductive. Nowadays, there is much controversy and indeed much skepticism on the part of those in the legal profession about modeling legal reasoning as inductive using the Bayesian calculus to attach probability values to statements and evaluate legal reasoning using conditional probability (Tillers 1989). In the meantime, argumentation-based technology has provided qualitative methods that can be used to identify, analyze, and evaluate arguments. In particular, argumentation schemes, standardized argument patterns different from the familiar deductive and inductive models of reasoning, have proved useful for this purpose (Wyner and Bench-Capon 2007). These developments have lent support to Wigmore's view that there is a science of proof underlying legal reasoning different from deductive logic (Sartor 2005; Prakken 2005, 2006). In this chapter it is shown how recent advances in argumentation show the value of modeling legal reasoning in this new way.

In this chapter, legal reasoning is divided into two broad categories: (1) the kind of reasoning that applies rules to cases and (2) the kind of reasoning used to determine what the facts of a case are. In this chapter, it is shown how to apply several centrally important argumentation schemes to a procedural sequence of case-based reasoning in which there is a successive refinement of cases. This sequence has an opening stage where the ultimate *probandum* is stated, an argumentation stage where arguments on both sides are put forward, and a closing stage. It will also be shown that it is necessary to distinguish between explanation and argument to better appreciate the role of explanation in legal reasoning. As will be shown, inference to the best explanation is another type of reasoning commonly used in legal argumentation and important for understanding how it works (Josephson and Josephson 1994).

D. Walton (✉)
University of Windsor, Centre for Research in Reasoning,
Argumentation and Rhetoric (CRRAR), Windsor, ON, Canada
e-mail: waltoncrrar@gmail.com

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_3

The first section studies the kind of reasoning that applies rules to cases in law, including the following forms of reasoning: argument from an established rule, argument from a verbal classification, and argument from precedent. By using an example from a Supreme Court case summary, it is shown how all three kinds of reasoning can be combined into a chain of reasoning that represents the structure of the evidence used to support an ultimate conclusion to be proved in a case. In this section, it is shown how rule-based reasoning of this kind is more complex than it might initially seem. One reason is that reasoning from precedent depends on argument from analogy. Another reason is that as rules are applied to cases and analogies are made from a precedent case to a given case, the rules need to be continually modified as they are re-applied to the series of cases. The second section introduces the argumentation scheme for argument from analogy, shows how it is based on a notion of similarity between pairs of cases, and shows how case-based reasoning is based on a chained sequence of similarity reasoning by analogy in which cases are successively refined over the continuing sequence. The third section discusses the distinction between reasoning and argument. The following forms of reasoning are analyzed and discussed: practical reasoning, value-based practical reasoning, reasoning from lack of evidence, abductive reasoning, and argument from perception. The fourth section extends the analysis to two forms of reasoning that draw from inferences from sources, argument from witness testimony and argument from expert opinion. The fifth section shows how the structure of reasoning exhibited in the first four sections is that of defeasible logic. The sixth section shows how the notion of proof, including the notions of standard of proof and burden of proof, needs to be defined within a procedural context of argumentation that has the three main stages stated above. The final section contains the conclusion.

1 Forms of Reasoning by Applying Rules to Cases

Legal reasoning is often visibly based on a form of inference called argument from an established rule in the argumentation literature (Walton, Reed and Macagno 2008, 343).

> *Major Premise*: If carrying out types of actions including the state of affairs *A* is the established rule for *a*, then (unless the case is an exception), *a* must carry out *A*.
>
> *Minor Premise*: Carrying out types of actions including state of affairs *A* is the established rule for *a*.
>
> *Conclusion*: Therefore *a* must carry out *A*.

In this form of reasoning, *a* is a rational agent that is capable of carrying out goal-directed actions and recognizing the consequences of its actions. An agent also has the capability of feedback, that is, the capability of changing its actions depending on their perceived consequences. This form of reasoning also contains the assumption that the agent has a knowledge base containing a set of established rules. The old idea of mechanical jurisprudence considers the application of rules to cases as a

straightforward application of deductive reasoning. The new approach of artificial intelligence and law sees the agent (judge or trier) as applying rules that can be defeated or overruled by exceptions as new evidence is introduced into its knowledge base.

Legal positivists, for example (Hart 1961), see law as consisting of two kinds of legal rules. The primary rules are the legal norms that regulate the activity of citizens and other persons. The secondary rules represent procedural norms that regulate the processes whereby the legislatures and courts put the primary legal rules into place and modify and apply them. But Hart recognized that both kinds of legal rules are inherently defeasible, meaning they admit of exceptions.

The term "defeasible" comes from medieval English contract law. It referred to a contract that has a clause in it that could defeat the contract in a case the circumstances of the case fit the clause. This meaning is now broadened to include the notion of a defeasible rule, a rule that is open to exceptions. Hart, in his famous paper "The Ascription of Responsibility and Rights" (1949, 1961), extended the usage of this term even further by writing about defeasible concepts. His most famous example is from *The Concept of Law* (1961). Consider the rule that no vehicles are allowed in the park. This rule could be defeated by special circumstances, for example during a parade, but it could also be defeated because of the open texture of the concept of a vehicle. Even though a car is classified as a vehicle, and would be excluded from the park, it may be debatable whether other objects such as a bicycle or a skateboard also fit into the same classification.

Consider the case of the drug-sniffing dog (Brewer 1996; Weinreb 2005). Suppose a trained dog sniffs luggage left in a public place and signals to the police that it contains drugs. Should this event be classified as a search according to the Fourth Amendment? If so, the evidence so obtained is not admissible as evidence. The problem is that the concept of a search is defeasible and law cannot define it by means of a set of necessary and sufficient conditions for closed to future revision because new cases may arise. Instead of providing closed essential definitions that give necessary and sufficient conditions, the best that can be done is to provide rules that may give necessary or sufficient conditions by indicating what types of things are included or excluded generally under the concept.

Weinreb (2005, 24) discussed two examples of such rules. One is the rule that if a police officer opens luggage and then observes something inside the luggage, the information collected is classified as a search. This is a narrow rule because it applies only to luggage, but it may still offer some helpful guidance in a case. The other is the rule that if a police officer obtains information about a person or thing in a public place without intrusion on the person or taking possession of or interfering with the use of the thing, it is not classified as a search. This rule is more general, but depends on the meaning of the prior concept of an intrusion on the person, as well as other concepts like that of taking possession of something or interfering with the use of something. These terms will, of course, also be open-textured and could possibly be subject to disputation. Questions arise quite often concerning how actions, events, and objects should be properly classified. In the example where the trained dog sniffs luggage left in a public place and signals to the police that it contains drugs,

the question is whether this event should be classified as a search according to the Fourth Amendment.

Arguing from a legal classification is a special form of reasoning in its own right. The following scheme represents argument from verbal classification (Walton, Reed and Macagno 2008, 319). Here, the constant *a* represents an individual that can be an object of any kind, including an event, a physical object, an animal, or a human being.

Individual Premise: *a* has property *F*.

Classification Premise: For all *x*, if x has property *F*, then *x* can be classified as having property *G*.

Conclusion: *a* has property *G*.

An ontology, a framework that specifies and organizes classes of concepts that can be used to represent the important features of cases (Ashley 2009, 8), can be brought forward to represent classifications of concepts to support legal reasoning about claims and issues. It includes representation of actual concepts like "animal," as well as legal concepts like "possession." In order to see how argumentation-based theories of legal reasoning in artificial intelligence would model the reasoning in a typical common law case, in addition to the scheme for argument from an established rule and the scheme for argument from verbal classification, we also need to take into account the working of a third scheme, called argument from precedent.

Most notably in common law countries, a ruling on a case is influenced by precedents. The most common type of argument from precedent used in legal reasoning applies to a current case and a prior case that has already been decided where the ruling can be applied to the current case (Schauer 1987). The argumentation scheme appropriate for this type of argument is the one for argument from precedent (Walton, Reed and Macagno 2008, 72).

Previous Case Premise: The source case is a previously decided case.

Previous Ruling Premise: In the source case, rule *R* was applied and produced finding *F*.

New Case Premise: The target case is a new case that has not yet been decided.

Similarity Premise: The target case is similar to the source case in relevant respects.

Conclusion: Rule *R* should be applied to the target case and produce finding *F*.

This way of configuring argument from precedent makes it a species of argument from analogy (Macagno and Walton 2009). In the next section, we will see how argument from classification is an extension of argument from analogy typically used in many arguments from precedent.

The following example can be used to illustrate how forms of reasoning like argument from an established rule and argument from a verbal classification can be used to form a chain of reasoning in a legal case that has the claim at issue in the case as its ultimate conclusion to be proved. In this US Supreme Court case (CSX Transportation, Inc. v. Alabama Department of Revenue et al. certiorari to the US Court of Appeals for the eleventh circuit No. 09-520, decided February 22, 2011) CSX claimed that the State of Alabama had discriminated against them

(http://www.supremecourt.gov/opinions/10pdf/09-520.pdf). The State taxes diesel fuel consumed by railroads but exempts interstate motor and water carriers. CSX claimed that this tax scheme discriminates against railroads in violation of the Railroad Revitalization and Regulatory Reform Act of 1976 which bars discriminatory taxation. The trial summary quoted below gives a concise account of the chain of reasoning used by the court to arrive at its decision.

> The key question thus becomes whether a tax might be said to "discriminate" against a railroad under subsection (b)(4) where the State has granted exemptions from the tax to other entities (here, the railroad's competitors). Because the statute does not define "discriminates," the Court again looks to the term's ordinary meaning, which is to fail to treat all persons equally when no reasonable distinction can be found between those favored and those not favored. To charge one group of taxpayers a 2% rate and another group a 4% rate, if the groups are the same in all relevant respects, is to discriminate against the latter. That discrimination continues if the favored group's rate goes down to 0%, which is all an exemption. To say that such a tax does not "discriminate" is to adopt a definition at odds with the word's natural meaning. This Court has repeatedly recognized that tax schemes with exemptions may be discriminatory. See, e.g., Davis v. Michigan Department of Treasury, 489 U. S. 803. And even Department of Revenue of Ore. v. ACF Industries, Inc., 510 U. S. 332, on which the Eleventh Circuit heavily relied in dismissing CSX's suit, made clear that tax exemptions "could be a variant of tax discrimination." Id., at 343. In addition, the statute's prohibition of discrimination applies regardless whether the favored entities are interstate or local. The distinctions drawn in the statute are not between interstate and local actors, as Alabama suggests, but between railroads and all other actors, whether interstate or local.

The central reasoning in this case can be visually represented as an inverted tree structure with the ultimate conclusion at the top, with the relevant arguments used to support that conclusion represented below in the argument diagram. The text boxes represent the statements that are the premises and conclusions of the arguments. The rounded nodes represent arguments, and the lines joining the nodes to text boxes represent inferences from premises to conclusions. In many instances, the conclusion of one argument becomes a premise in the next one, producing a chain of argumentation terminating in the root of the tree at the top.

We can see that in some instances an argument is represented as having one premise, while in other instances an argument has multiple premises. In some cases, we have two or more separate arguments going to support the same conclusion. For example, the two arguments for established rule just under the conclusion that attacks might be said to discriminate against a railroad under subsection, etc., each independently support it. This type of structure is sometimes called a convergent argument. Where we have two or more premises that are combined together to support the same conclusion so that each premise needs the others, this configuration is called a linked argument. The argument from the rule of definition at the top is a case in point.

To keep the argument diagram in Fig. 1 as clear and concise as possible at this point, no implicit premises have been represented. For example, three arguments just under the main conclusion have implicit premises, but these premises are not represented on the diagram. Argumentation schemes representing typical forms of defeasible reasoning are placed in all of the argument nodes in the diagram except one.

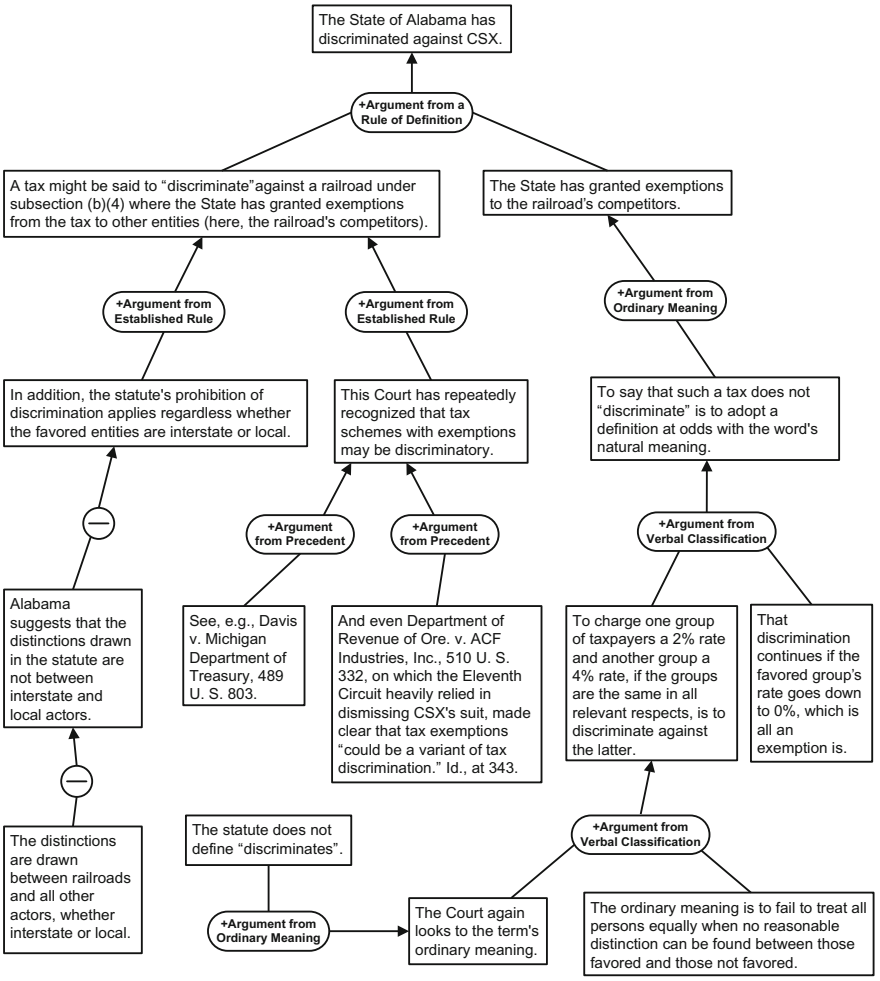


Fig. 1 Chain of argumentation in the CSX case

Not all instances of argumentation have a known argumentation scheme that represents the type of argument, and its form, that enables the inference transition from the premises to conclusion. All the argumentation schemes in Fig. 1 are explained later in the chapter (except one, called "argument from ordinary meaning"). Pro-arguments, arguments supporting the conclusion, are indicated by a plus sign in the argument node. Contra-arguments, arguments against a conclusion, are indicated by a minus sign in the argument node. This argument map was drawn using the Carneades Argumentation System, a formal argumentation system that has an argument mapping tool specially designed to represent legal argumentation. How the system works will be briefly explained later in part four of the paper.

Representing the reasoning used in this case presented in the quoted Supreme Court summary above offers a nice way of visually grasping how legal reasoning works by applying rules to cases and by a process of classifying key terms. Once we realize that this kind of reasoning is defeasible and that it is based on terms that are open-textured and very susceptible to argumentation by the opposed side, it no longer seems as simple or straightforward as it might have been when we call it "applying rules to cases." Understanding how rules should be applied to cases takes us to the subject of reasoning from precedent cases.

Reasoning from a precedent depends on an underlying form of reasoning from analogy based on the similarity of the source case to the target case. On this model, rules are continually being modified as they are applied over and over again to a series of cases. By means of examining a number of examples of legal rulings that demonstrate how actual legal method in the common law systems works from examples that result in changes of legal rulings in successive trials, Levi (1949, 104) was led to the conclusion that "legal reasoning has a logic of its own." These examples revealed a contrast between "logic and actual legal method" (104). According to Levi's analyses of how legal decisions were arrived at in the extensive examples he provided, particular entities are classified as falling under general terms that occur in rules that are applied to cases and then modified when the new case is decided on in a different way. According to Levi (1949, 8), this process of legal reasoning has three stages. The first stage is the creation of a legal concept built up from cases. The second stage continues this process of reasoning by example by fixing the concept. The third stage is the breakdown of the concept. The example given by Levi (1949, 14) is the "inherently dangerous rule." In commercial transactions where one party sells something to another party and the second party is injured by using the product, differences in liability turn around the issue of whether a commercial product can be classified as "inherently dangerous" or not. This category was gradually expanded through a series of cases where one product was judged to be similar enough to another product that had already been classified as inherently dangerous so that the second product could also be placed in the same category.

2 Case-Based Reasoning from Analogy

The literature on argument from analogy in fields spanning logic, argumentation studies, computer science, and law, is enormous. Many proposals have been put forward to represent argument from analogy as a form of reasoning or argumentation scheme, and there is no space to try to summarize them here. We can only refer the reader to the multi-disciplinary bibliography of Guarini et al. (2009). However, the use of argument from analogy in case-based reasoning (Ashley 1988) is central. Here, we concentrate on two particular proposals to represent the structure of this argumentation scheme that provides a useful contrast to focus the discussion.

The simplest argumentation scheme for argument from analogy can be represented by this first version from (Walton, Reed and Macagno 2008, 315).

Similarity Premise: Generally, the source case is similar to the target case.

Base Premise: *A* is true (false) in the source case.

Conclusion: *A* is true (false) in the target case.

Let us call this scheme the basic scheme for argument from analogy. The assumption behind the basic scheme for argument from analogy is that there exists a similarity between two cases where *A* holds in the source case and can shift a weight of evidence to make it plausible that *A* holds in both cases. This kind of argument is defeasible, and it can in some instances even be misleading and fallacious, as the traditions of informal fallacies warn us (Hamblin 1970). It is an important kind of argument to study, because so much of our reasoning is based on it (Schauer 2009). But how can similarity be modeled in such a way that we have evidence that is useful to determine whether and how the source case is similar to the target case? An example is helpful.

Barry Bonds hit a valuable home run ball into the stands in the case of Popov v. Hayashi, a trial concerning the issue of which one of two fans could claim ownership rights to the ball. The reasoning in the trial partly turned on some historical precedent cases that concerned the hunting and fishing of wild animals. The trial became a classic example for study on how case-based reasoning can be applied from similar precedent cases to an analogous case at issue (Wyner et al. 2007; Bench-Capon 2009, 2012). The ball went into the upper portion of the webbing of a glove worn by a fan, Alex Popov, but as it entered his glove, he was thrown to the ground by a mob of fans trying to get the ball. While Popov was pinned to ground by the mob, a nearby fan, Patrick Hayashi, not part of the mob that had knocked Popov down, pocketed the loose ball. When the man making a videotape pointed the camera at Hayashi, he held the ball in the air for the others to see. Hayashi was not at fault for the assault on Popov. According to generally accepted rules of baseball, a fan has the right to keep the baseball he has caught in the stands. But such a catch only bestows this right when the fan has the ball in his hand or glove and the ball remains there after its momentum has ceased. If no one catches the baseball, any person in the stands may come to own it by picking it up. According to these accepted rules, it would appear that Hayashi had the right to ownership of the ball. However, Popov also claimed this right and took the case to court.

The contested issue in the trial that took place in the Superior Court of California City and County of San Francisco was about which party should properly be said to have possession of the ball, but the outcome could not be decided by simply applying the legal concept of possession (McCarthy 2002, 5). Some comparable historical precedent cases were presented where there was pursuit of an animal that the pursuer failed to catch because somebody or something intervened, and the issue was whether the pursuer could claim possession of the animal. In the case of Pierson v. Post (3Cai. R. 175; 1805 N.Y. LEXIS 311), Pierson was chasing a fox with hounds when Post captured and killed it, even though he was aware that it was being pursued. The court decided in favor of Post on the basis that mere pursuit did not give Pierson a right to the fox. In the case of Young v. Hitchens (6 Q.B.606 (1844)), Young was a fisherman who spread his net, but when it was almost closed, Hitchens went through the gap with a net and caught the fish. The court decided in favor of Hitchens.

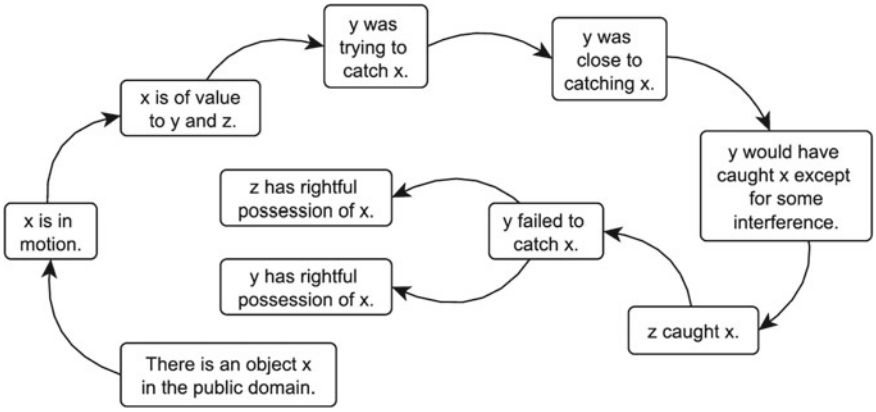


Fig. 2 Basis of similarity of the animal cases to the Popov case

In Keeble v. Hickeringill ((1707) 103 ER 1127), *P* owned a pond and made his living by shooting ducks lured onto it with decoys and selling them. *D* used guns to frighten the ducks away from the pond. This case was decided in favor of *P*. In Ghen v. Rich (8 F.159 D. Mass, 1881), Ghen harpooned a whale from his ship, but it was washed ashore found by another man who sold it to Rich. The generally accepted rule of whaling was that the party who finds the whale should report it and can then collect a fee. This case was decided in favor of Ghen.

In the end, after examining many arguments, Judge McCarthy (2002, 9) ruled that any award to one party would be unfair to the other and that each had an equal and undivided interest in the ball. The historical precedent cases are nevertheless interesting in their own right as part of the evidence in the case, because they raise questions about what is meant by "similarity" when a precedent case is seen to be analogous to a case at issue. In many respects, these wild animal cases are not similar to the baseball case at all. As Gray (2002, 1) observed, "a baseball at the end of its arc of descent is not at all like a fox racing across the commons, acting under its own volition, desperately attempting to evade death at the hands of its pursuers." But there is a general similarity underlying the pattern of action in all these cases. In general, they are all about an agent trying to catch something to possess it and about some kind of interference that prevented him from obtaining it, leading to a question of who has the right to possess it (Walton 2010). The similarity can be visualized as an abstract sequence of actions and events that hangs together in a pattern shown in Fig. 2, where *x* is a variable for an object and *y* and *z* are variables for agents. The open arrows represent transitions from one action or event to another in a story scheme (Pennington and Hastie 1993; Bex 2009), while the arrows in the middle represent inferences drawn to a conclusion.

On this view, legal reasoning is a sequence of steps based on similarity between pairs of cases in which a rule applied to one case can also be applied to a second case

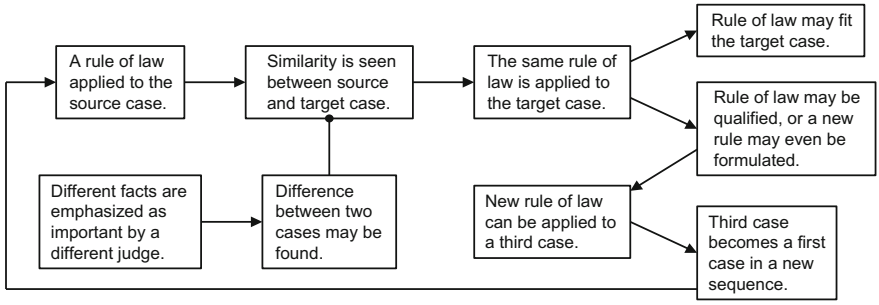


Fig. 3 Sequence of case-based similarity reasoning from analogy

that is taken to be similar to the first one. Essentially, the sequence of reasoning is based on argument from analogy.

But that is not the end of the sequence. There are two possibilities. The argument from analogy may be defeated when a significant difference between the two cases is found (Ashley 2006). Such a difference arises because different facts are emphasized as important by a different judge. The other possibility is that the argument from analogy may be successfully applied to the second case, and this in turn may have two possible outcomes, shown in Fig. 3.

One outcome is that the rule may fit the second case, and in this instance, the same conclusion will be drawn in the second case as was drawn in the first case. The other is that the rule will be changed, by being qualified or otherwise modified, or it may even by being reconfigured by a different rule that replaces the earlier rule. When this outcome occurs, the new rule can be applied to a new case, the third case in the sequence, which starts the whole process over again, or establishes the new rule, and the new third case is applied to a fourth case on the basis of a similarity seen between the third and fourth cases.

Levi (1949, 3) sums up the process by not agreeing that it is a system of applying rules to facts, but rather by showing that it is a complex procedure in which "the rules are discovered in the process of determining similarity or difference." This kind of reasoning is dynamic, because there is a continual feedback process in which the new cases may lead to rulings that are inconsistent with the rules held in the old cases, in which the terms used in the new cases may result in different classifications from those in the old cases.

This process cycles on and on indefinitely as the rules of law are refined, creating precedents by being applied to new cases, as shown in Fig. 4.

The sequence of reasoning shown in Fig. 4 is a process of successive refinement based on analogies between cases. The best model we have of this procedure of reasoning to conclusions is that of case-based reasoning, a technology used to solve a problem posed in a given case by drawing on similar cases retrieved from a database of past cases (Ashley 2006). The solution to the problem posed in the given case is achieved by matching the given case against the retrieved cases by a process of analogy that selects similar cases. The HYPO system produces point–counterpoint

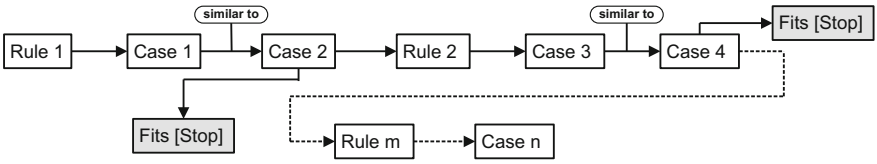


Fig. 4 Successive refinement of cases as a continuing sequence

arguments in trade secrets law, and the CATO system teaches law students how to create case-based arguments. HYPO analyzes a given case by retrieving similar cases from its knowledge base, and makes a judgment concerning which cases are most "on point." CATO has templates for argument moves, such as the one for argument from analogy, and rules that show how to attack a rule (Ashley and Rissland 2003, 41). It can generate an argument for one side while also producing counterarguments that support the other side. Case-based reasoning requires argumentation schemes, especially argument from analogy and argument from precedent (Ashley 2009), and combines these arguments into chains or trees of argumentation modeled on the chain of reasoning shown in Fig. 4.

3 Reasoning and Argument

From a logical, as opposed to a psychological point of view, reasoning may be defined as a series of steps of inference in which some propositions are inferred from others (Walton 1990a, b, 404). Reasoning is sequential and best visualized abstractly as an argument diagram where propositions are contained in text boxes. These text boxes are joined to other text boxes with arrows representing inferences from some propositions to others. Although it is possible to have (as in multi-conclusion logic) premises that lead by inferences to two or more distinct conclusions, it simplifies our view of reasoning if we exclude this possibility. On this more restricted view, reasoning can have several starting points (premises) but the inference drawn from them always leads to a single conclusion. On this view, however, reasoning can be sequential. That is, some premises can lead to a particular conclusion, and this conclusion can then become a premise in the next step of inference to a different conclusion. Such a configuration is called chaining of reasoning in artificial intelligence. If we look at an argument this way, its structure can often be represented as a type of tree referred to as a tree with a single root proposition as the ultimate conclusion. This tree branches outward to a series of connected inferences all leading away from the ultimate conclusion that is the root. An excellent example is the reasoning used in a Supreme Court case represented as an argument diagram using the Araucaria tool for argument visualization (Fig. 1). The reader can see that the diagram of the chain of reasoning as an inverted tree structure with one proposition designated as the ultimate probandum represented as the root of the tree.

To help clarify this definition of reasoning, it is helpful to draw a distinction between epistemic reasoning and practical reasoning. Epistemic reasoning is used to determine whether a proposition is true or false, or whether it is unknown to be either true or false, based on the knowledge of an agent. The simplest form of practical reasoning, called instrumental practical reasoning, has the following form (Walton et al. 2008, 323). The first-person pronoun "I" represents a rational agent that has goals, some (though possibly incomplete) knowledge of its circumstances, the capability of altering those circumstances, and the capability of perceiving the consequences of so acting.

Major Premise: I have a goal *G*.

Minor Premise: Carrying out this action *A* is a means to realize *G*.

Conclusion: Therefore, I ought (practically speaking) to carry out this action *A*.

Practical reasoning can move forward, from a goal to an action, as part of agent-based deliberation, but it can also be used backward by inference to the best explanation (see below) to reconstruct an agent's internal mental states such as motive or intent, based on an agent's known actions and words.

Practical reasoning can be undercut by citing possible negative consequences of the proposed action. Such a form of attack is a species of reasoning in its own right (Walton et al. 2008, 332).

Premise: If *A* is brought about, then bad consequences will occur.

Conclusion: *A* should not be brought about.

By its use of the word "bad," this form of reasoning is seen to be based on values that the agent may be presumed to have and hence it is a species of value-based reasoning (Bench-Capon 2003). However, arguments from positive or negative values can also operate as individual arguments in their own right (Bench-Capon 2003) independent of argument from consequences. The first argumentation scheme represents the argument from positive value.

Major Premise: If value *V* is positive, it supports commitment to goal *G*.

Minor Premise: Value *V* is positive as judged by agent *a*.

Conclusion: *V* is a reason for *a* to commit to goal G.

The negative counterpart is called argument from negative value.

Major Premise: If value *V* is negative, it attacks commitment to goal *G*.

Minor Premise: Value *V* is negative as judged by agent *a*.

Conclusion: *V* is a reason for *a* to retract commitment to goal G.

Argument from values is combined with instrumental practical reasoning to yield the scheme for value-based practical reasoning. This scheme was first formulated in the following form by Atkinson and Bench-Capon (2007, 861).

Circumstances Premise: S_1 is the case in the current circumstances.

Action Premise: Performing *A* in S_1 would bring about S_2.

Goal Premise: G would be realized in S_2

Value Premise: Achieving the goal G would promote the value V.

Conclusion: Action A should be performed.

The following critical questions match the scheme for value-based practical reasoning.

CQ_1: Is V is a legitimate value?

CQ_2: Is G is a worthy goal?

CQ_3: Is action A possible?

CQ_4: Does there exist an action that would bring about S_1 more effectively than A?

CQ_5: Does there exist an action that would realize the goal G more effectively than A?

CQ_6: Does there exist an action that would promote the value V more effectively than A?

CQ_7: Would performing A in S_1 have side effects which demote V or some other value?

An example of value-based practical reasoning (Atkinson et al. 2006, 82) is: I may diet to lose weight, with the goal of not being overweight, to promote the value of health. In the value-based scheme, the notion of a goal is separated into three elements: the state of affairs brought about by the action, the goal (the desired features in that state of affairs), and the value. The value is defined as the reason why those features are desirable. The structure is based on an Action-based Alternating Transition System (Wooldridge and van der Hoek 2005) in which an agent performs an action by moving from a current state of affairs to a new one with many differences that may make the new state of affairs better with respect to some value of the agent.

Practical reasoning is used in a situation of uncertainty and incomplete knowledge where an agent has to make a decision in a given situation that is constantly changing, based on its goals and its knowledge of that situation. This assumption that there is new incoming information to the agent because the situation is constantly changing is called the open world assumption in artificial intelligence. The conclusion is whether a particular course of action should be taken or not. Sometimes doing nothing (inaction) needs to be represented as a possible course of action, because doing nothing at all can often have negative consequences and can affect the agent's goals. Although the open world assumption is typical of practical reasoning of the kind that takes place in realistic decision-making, it is also possible in some instances to invoke what is called the closed world assumption. The closed world assumption rules that no further evidence will count as relevant because the knowledge already available can be regarded as exhaustive of all the relevant evidence for the conclusion.

In a common law criminal trial, the presumption of innocence is taken to shift the burden of proof onto the prosecution to prove its claim of guilt to the standard of beyond a reasonable doubt. All the defense has to do is to cast doubt on the argumentation put forward by the prosecution. This asymmetrical management of the burden of proof in a trial is evocative of the argument from lack of evidence, often called the argument from ignorance in the literature on informal fallacies in logic. However, when used in this context, the following version of the argument is reasonable: It has not been proved that the defendant is guilty; therefore, the defendant

should be presumed to be innocent (not guilty). This kind of reasoning is reasonable, except for a complication in cases of Scottish jurisdiction, where the jury can return a verdict of not proven. The question of how burden of proof should be determined in different settings of argumentation is taken up in Sect. 6.

Invoking the closed world assumption has been identified as a form of epistemic reasoning called argument from lack of evidence (Walton, Reed and Macagno 2008, 327).

> *Major Premise*: If *A* were true, then *A* would be known to be true.
>
> *Minor Premise*: It is not the case that *A* is known to be true.
>
> *Conclusion*: *A* is not true.

This form of reasoning depends on how far along the search for evidence has progressed in a given case and may therefore be regarded as defeasible, unless the knowledge base can be closed on the grounds that all the available evidence has been collected and processed. This ground is called the standard of proof (see Sect. 6). The following inference is an example of a typical instance of this kind of reasoning in history.

> *Minor Premise*: There are no known instances of Romans being awarded medals for bravery in battle posthumously.
>
> *Major Premise*: If there were instances of Romans being awarded medals for bravery in battle posthumously, we would know of them.
>
> *Conclusion*: Therefore, the Romans did not award medals for bravery in battle posthumously.

To support the minor premise, the following statements might be offered as evidence: We would see evidence on tombstones or in written records of battles. Sometimes reasoning from lack of evidence can be provisionally acceptable, depending on the standard of proof, even though it is based on negative evidence.

If the closed world assumption cannot properly be made, there is the possibility of new information that can affect the outcome of practical reasoning. It is especially important for adequate practical reasoning that a rational agent be open to new incoming information and be flexible in taking this information into account in modifying its goals and actions accordingly.

Another form of inference that is very important in legal reasoning is abductive reasoning, or inference to the best explanation (Pardo and Allen 2008). According to Josephson and Josephson (1994, 14), abductive inference has the following form, showing its structure as inference to the best explanation. *H* is a hypothesis.

- *D* is a collection of data.
- *H* explains *D*.
- No other hypothesis can explain *D* as well as *H* does.
- Therefore, *H* is probably true.

An example quoted from (Wigmore 1940, 420) shows how he analyzed cases of legal evidence as instances of inference to the best explanation.

> The fact that *a* before a robbery had no money, but after had a large sum, is offered to indicate that he by robbery became possessed of the large sum of money. There are several other possible explanations - the receipt of a legacy, the payment of a debt, the winning of a gambling game, and the like. Nevertheless, the desired explanation rises, among other explanations, to a fair degree of plausibility, and the evidence is received.

The evidence put forward in this example has the form of an inference to the best explanation where the conclusion was arrived at by means of a choice among several competing explanations of given facts.

Another important form of reasoning in law is the drawing of an inference based on perception (Pollock 1995, 41).

> *Premise 1*: Person *P* has a φ image (an image of a perceptible property).
>
> *Premise 2*: To have a φ image (an image of a perceptible property) is a *prima facie* reason to believe that the circumstances exemplify φ.
>
> *Conclusion*: φ is the case.

Pollock (1995, 41) offered the following argument as an example.

> *Minor Premise*: This object looks red to me.
>
> *Major Premise*: When an object looks red, then (normally, but subject to exceptions) it is red.
>
> *Conclusion*: This object is red.

This argument is defeasible, as Pollock pointed out, since even objects that are not red can look red when illuminated by a red light. It is a species of defeasible reasoning that can give a reason to accept its conclusion, provided there is no reason to think that the situation is exceptional.

The most important forms of legal reasoning in law are defeasible (Hart 1949, 1961). Other good examples are instances of drawing inferences from sources, like reasoning from witness testimony and expert opinion. Such forms of reasoning are not well modeled by a deductive logic based on the major premise that what an expert says is always true. Fitting reasoning from expert opinion testimony into a deductive model would render it into a fallacious form of reasoning by making it intolerably rigid. Instead, we need to look to defeasible reasoning.

4 Reasoning by Drawing Inferences from Sources

Drawing an inference from perception clearly represents a kind of reasoning we utilize all the time, not only in law but also in scientific reasoning and in everyday conversational reasoning. As skeptics have often noted, this kind of reasoning is defeasible. Unfortunately, however, in many situations where we have to draw a conclusion on what to do or what to accept as a hypothesis, the reasoner himself has no direct access to data through perception. In such cases, we have to rely on information derived from sources. Source-based reasoning is vitally important in law, because many of the supposed facts happened in the past. One of the most important

forms of source-based reasoning and law is inference from witness testimony. The argumentation scheme for this form of reasoning is given in Walton (2008, 60). It has three premises.

Position to Know Premise: Witness *W* is in a position to know whether *A* is true or not.

Truth-Telling Premise: Witness *W* is telling the truth (as *W* knows it).

Statement Premise: Witness *W* states that *A* is true (false).

Conclusion: Therefore (defeasibly) *A* is true (false).

In argument from witness testimony, two key premises are the position to know premise and the truth-telling premise. It is assumed that the witness is basing what she says on her genuine knowledge of some real situation or true set of facts. Moreover, it is assumed that she is telling the truth about those facts, as she saw or knows what she witnessed. These assumptions pose some constraints on witness testimony as a form of argument. It needs to be assumed that the account the witness has presented is internally consistent, and is consistent with known facts of the case that can be verified by independent objective evidence. These matters can be tested by using the following critical questions (Walton 2008, 61) matching the scheme for argument from witness testimony.

Internal Consistency Question: Is what the witness said internally consistent?

Factual Consistency Question: Is what the witness said consistent with the known facts of the case (based on evidence apart from what the witness testified to)?

Consistency with Other Witnesses Question: Is what the witness said consistent with what other witnesses have (independently) testified to?

Trustworthiness Question: Is the witness personally reliable as a source?

Plausibility Question: How plausible is the statement A asserted by the witness?

Bias Question: Is there a bias that can be attributed to the account given by the witness?

If the witness was really in a position to know the facts of a case and is giving an honest and accurate report of what she saw or heard, this should produce an account that is internally inconsistent. Or if it does not appear to be consistent in some points, the apparent inconsistency should be able to be explained or resolved. But consistency can only be tested by probing into the account given by the witness and seeing if her story "stands up" under questioning during examination. This procedure is analyzed in the next section.

Another form of argument that is important in legal reasoning is that of argument from expert opinion. Epistemic reasoning to a conclusion based on expert opinion testimony as an admissible form of evidence requires that the source be qualifiable as an expert. For example, ballistics experts and DNA experts are often used to give expert testimony as evidence in trials, but they must qualify as experts. The most basic version of the form of reasoning from expert opinion is modified from the one in Walton Reed and Macagno (2008, 310) as follows.

Major Premise: Source *E* is an expert in field *F*.

Minor Premise: *E* asserts that proposition *A* is true (false).

Second Minor Premise: *A* is within *F*.

Conclusion: *A* is true (false).

It is not helpful to treat this form of reasoning as deductive, for that would amount to taking an expert as an infallible source of knowledge. Taking that approach makes argumentation susceptible to many serious problems known to be associated with the fallacious misuse of argument from expert opinion. According to the contrasting approach of Walton (1997, 223), an argument from expert opinion should be evaluated by the asking of six basic critical questions.

Expertise Question: How credible is *E* as an expert source?

Field Question: Is *E* an expert in the field *F* that *A* is in?

Opinion Question: What did *E* assert that implies *A*?

Trustworthiness Question: Is *E* personally reliable as a source?

Consistency Question: Is *A* consistent with what other experts assert?

Backup Evidence Question: Is *E*'s assertion based on evidence?

According to Walton (1997), if a respondent asks any one of the six critical questions, the original argument defaults until the question has been answered adequately.

A problem with using critical questions to evaluate cases where expert opinion is used as a source of evidence is that we can no longer use an argument diagram to summarize, analyze or evaluate the basic evidence in a case and display its structure as a sequence of reasoning. The reason is that everything that appears in the text box on a standard argument diagram needs to be a statement, a proposition that is either true or false. It is harder to analyze the structure of questions, even though they are certainly very important as devices in both everyday and legal argumentation, for example in examining a witness. Using critical questions definitely takes us outside the realm of reasoning to the realm of argument, where claims are made and subjected to doubt by the asking of critical questions by an opponent.

We can see the problem more explicitly if we ask what happens when a critic asks a critical question. If the critic asks the opinion question, in other words if he asks the arguer who has appealed to argument from expert opinion to quote the specific statement that the expert made that supposedly implies *A*, then the arguer certainly has to respond to this reasonable question by presenting the critic with a specific proposition, for example by quoting exactly what the experts said. If the arguer fails to carry out this reasonable kind of request, this argument from expert opinion will surely fail to be persuasive. However, suppose the critic asks the trustworthiness question: is *E* personally reliable as a source? It could be perfectly reasonable for the arguer to shift the burden of disproof back to the questioner by replying, "of course she is personally reliable, for after all she is an expert, and if you wish to make any allegation to the effect that she is not personally reliable, you had better prove that." This kind of reply would certainly be sufficient as an adequate answer to the question. In other words, there are two kinds of critical questions. When the first kind of critical question is asked, any failure to answer inappropriately will defeat the argument. When the second kind of critical question is asked, the burden shifts from the arguer to the questioner to support the critical question with further argument. Because of these crucial differences between the critical questions, it

seems impossible to represent them in any straightforward way as statements that are additional assumptions of the argument.

Fortunately, however, there is a way that we can represent critical questions on an argument diagram by treating them as additional premises that need to be added to the given premises in the argumentation scheme (Walton and Gordon 2005). Artificial intelligence has found a way to do this by using the Carneades system (Gordon 2010). Carneades is a mathematical and computational model that defines mathematical properties of arguments that are used to identify, analyze, and visualize real arguments. By applying argumentation schemes, Carneades analyzes and evaluates the acceptability of arguments, based on proof standards, for example preponderance of the evidence. Carneades takes the approach that the way critical questions are modeled depends on the individual argumentation scheme, by distinguishing three kinds of premises. Ordinary premises are just the regular premises of an argumentation scheme that are explicitly given in the scheme itself. But there are two additional kinds of premises not stated in the scheme. Assumptions are to be acceptable unless called into question. Exceptions are modeled as premises that are not assumed to be acceptable and that can defeat an argument as it proceeds. Ordinary premises of an argument, like assumptions, are assumed to be acceptable, but they must be supported by further arguments in order to be judged acceptable.

In Fig. 5 the conclusion of the argument from expert opinion is represented by the text box at the far left stating that *A* is true. The argument from expert opinion is represented in the node with a plus sign in front of its name, indicating it is a pro-argument supporting the conclusion that *A* is true. In the list of premises on the right of the node, the first four are the ordinary premises of the scheme for argument from expert opinion. Hence, they are labeled as being in assumptions, statements that are assumed to hold, but are subject to critical questioning. Since there are no arguments against them, or critical questions directed to them, they are shown as acceptable in Fig. 5 by darkening the text boxes in which they appear. The next two premises are also classified as assumptions. Asking a critical question challenging either one of these for assumptions will shift the burden of proof onto the proponent of the argument from expert opinion and temporarily defeat it until the proponent replies to the question appropriately. Since they have not been questioned or attacked either, they are also shown as accepted. The two critical questions at the bottom, the trustworthy question and the "consistency with other what experts say" question, are represented as exceptions. These two critical questions are shown on the diagram as undercutters, because they are displayed as contra-arguments, indicated by the minus signs in their nodes. An undercutter is an argument that attacks an argument node rather than attacking a premise or conclusion (a proposition).

The "consistency with other what experts say" question is an undercutter with a premise displayed in white text box indicating that it is not accepted. Notice however that the other one of these premises, the trustworthiness premise, has an argument from bias supporting the statement that the expert is not trustworthy. Since this undercutter has been backed up by evidence to support it, in the form of the argument from bias, it successfully defeats the argument from expert opinion. Hence, despite the fact that the other four premises above it are assumed to hold, the argument

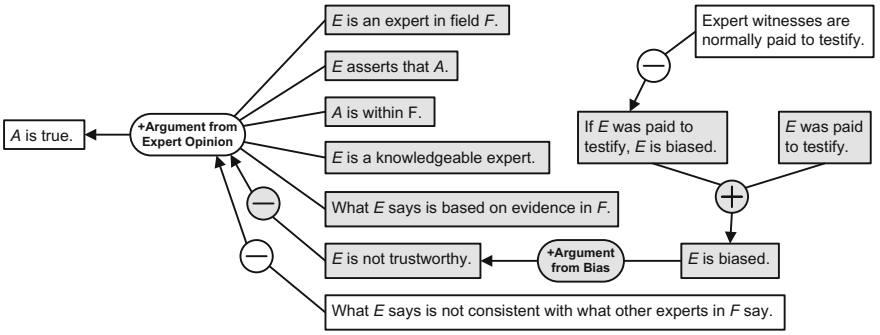


Fig. 5 Argument from expert opinion with undercutter in a Carneades Map

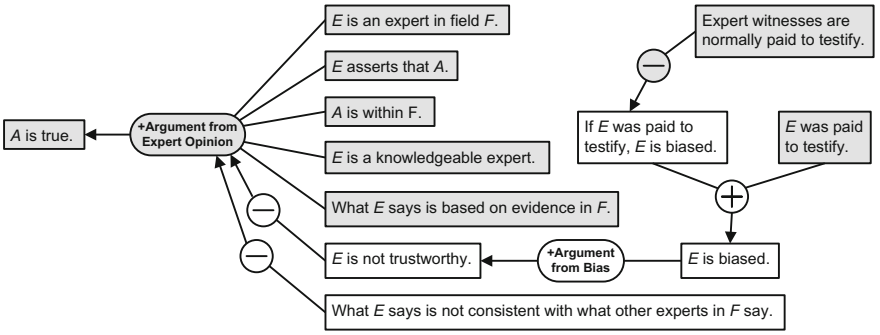


Fig. 6 Counterattack to the undercutter in an argument from expert opinion

from expert opinion fails. Notice however that one of the premises of the argument supporting the claim that *E* is biased is attacked by a counterargument stating that expert witnesses are normally paid to testify. In other words, the argument is that since expert witnesses are normally paid to testify, if a particular expert is paid to testify, that should not be taken as evidence that he is biased. However, the way the argumentation is represented in Fig. 5 to proposition that expert witnesses are normally paid to testify is not accepted, nor is it supported by a further argument.

But what would happen if the proposition that expert witnesses are normally paid to testify were to be accepted? This new evidential situation is shown in Fig. 6.

Notice at the top right of Fig. 6 a contra-argument based on the premise that expert witnesses are normally paid to testify has been used to attack one of the premises of the argument from bias. On the basis of this attack, one of the premises supporting the conclusion that *E* is biased is no longer accepted. Hence, the argument from bias is defeated by this premise attack, and the topmost undercutter now fails to defeat the argument from expert opinion. Since there is no supporting evidence behind the other undercutter displayed in the text box at the bottom of Fig. 6, the argument from expert opinion is restored. In this situation, Carneades will automatically display the ultimate conclusion that is true as "accepted," once the user has input the information

that all the other premises of the argument from expert opinion are assumed to hold, as shown in Fig. 6. So we can see that Carneades is a dynamic system that takes into account new arguments brought forward from its knowledge base, so that in some instances an argument that was formerly used to successfully refute an argument can be defeated by a new argument.

We have shown how Carneades incorporates defeasible logic and builds on it to provide a computational tool that not only enables us to do argument mapping, but to represent the critical questions matching defeasible argumentation scheme on an argument map, and to track new relevant evidence coming in from its knowledge base.

5 Defeasible Logic

Defeasible logic is a logical system (Nute 1994) that models reasoning used to derive provisional conclusions from partial and sometimes conflicting information. Using this kind of reasoning, a conclusion can be tentatively accepted, subject to new evidence that may come to be known at some point in the future of an investigation. This new evidence may require the retraction of the conclusion that was formerly accepted, based on the evidence that was available at that earlier point. The use of this kind of defeasible reasoning is highly appropriate during an investigation or sequence of argumentation where the inflow of new evidence may be closed off, even though it may later be reopened, for example in an appeal. Once the investigation has been closed off, the conclusion of the reasoning can be accepted as final, for the purposes of the investigation. But before that point, where there is argumentation on both sides, and all the arguments are not in, the reasoning needs to be regarded as defeasible.

The basic units of defeasible logic are called facts and rules. Facts are statements that are accepted as true within the confines of a discussion. Statements, also called propositions, are denoted by letters, A, B, C, …, using subscripts if necessary. There are two kinds of rules in defeasible logic, called strict rules and defeasible rules. Strict rules are absolutely universal in the sense that they do not admit of exceptions. An example of a strict rule would be the universal generalization "All penguins are birds." In defeasible logic, a strict rule has the form of a material conditional with a conjunctive antecedent of the following form: $A_1, A_2, A_n, \ldots, \rightarrow B$. In this kind of conditional, it is not possible for all the A_i to be true and the B false. Defeasible rules are rules that are subject to exceptions, and that may fail if an exception is shown to exist in a given case. An example of a defeasible rule would be the statement "Birds fly," meaning that birds generally fly or that birds normally fly, but not implying that all birds fly without any exception being allowed. A defeasible rule has the form A_1, A_2, A_n …, => B, where each of the A_i is called a prerequisite. The set of all the A_i taken together is called the antecedent, and the statement B is called the consequent. For example, suppose that Tweety is a bird, but we also know that he is a penguin, and that we know that penguins cannot fly. In light of our knowledge of this exception, the

conclusion that Tweety flies cannot be inferred and has to be retracted. The rule still holds, but the inference itself fails to support the conclusion any longer. Defeasible logic is the best way to represent the structure of reasoning of argumentation schemes of the kind most commonly used in law, for as shown above, this type of reasoning is defeasible.

It is a problem that the forms of reasoning that we are familiar with from deductive logic and from the kind of inductive logical reasoning used in the Bayesian rules do not appear to fit argumentation schemes of the defeasible kind illustrated above. However, Verheij (2001, 232) showed that these defeasible argumentation schemes fit a form of argument he called *modus non excipiens*, which has the following form.

As a rule, if *P* then *Q*

P

It is not the case that there is an exception to the rule that if *P* then *Q*

Therefore *Q*

Verheij showed how this form of argument can be used for evaluating defeasible inferences like the Tweety argument: If Tweety is a bird, Tweety flies; Tweety is a bird; therefore, Tweety flies. This form of argument was called defeasible *modus ponens* (DMP) by Walton (2002). A version of an example from (Copi and Cohen 1998, 363) can be used to illustrate DMP: If he has a very good defense lawyer, he will be acquitted; Bob has a very good defense lawyer; therefore, he will (likely) be acquitted. This argument is clearly defeasible, for even though Bob has a good lawyer, he may not be acquitted.

Using a concept from defeasible logic called defeasible implication, or the defeasible conditional, represented by the symbol =>, we can represent DMP is having the following form.

Major Premise: *A* => *B*

Minor Premise: *A*

Conclusion: *B*

The first premise is the conditional, "If *A* is true then generally, but subject to exceptions, *B* is true." In some instances, the argumentation schemes above, for example the argument from negative evidence, explicitly have the DMP form. In other instances, the scheme for argument from expert opinion for example, the scheme can be put into the DMP form by adding an implicit premise as an additional assumption.

To see how this works using an example, let us consider the following expanded version of the argument from expert opinion scheme.

Major Premise: Source *E* is an expert in subject domain *S* containing proposition *A*.

Minor Premise: *E* asserts that proposition *A* (in domain *S*) is true (false).

Conditional Premise: If source *E* is an expert in a subject domain *S* containing proposition *A*, and *E* asserts that proposition *A* is true (false), then *A* may plausibly be taken to be true (false).

Conclusion: *A* may plausibly be taken to be true (false).

If you look at this version of the scheme, you can see that the argument from expert opinion has a *defeasible modus ponens* structure as an inference.

Major Premise: (*E* is an expert and *E* says that *A*) => A

Minor Premise: *E* is an expert and *E* says that *A*

Conclusion: *A*

This form of argument is not exactly the same as DMP because the conditional in the major premise has a conjunctive antecedent. The scheme has this form: (*A* & *B*) => *C*, *A* & *B*, therefore *C*. Nevertheless, it is a substitution instance of the DMP form. It is fair to say that in its general outline it has the structure of the DMP form of inference.

The analysis so far, however, does not take into account the critical questions for the argument from expert opinion. There was a suggestion made by (Reed and Walton 2003, 202) that the conditional premise could be expanded to take the critical questions into account in a still more fully expanded version of the scheme. This proposal is now easily carried out using the Carneades system of treating the critical questions as additional assumptions or exceptions in the scheme. This form of the argument can now be seen as fitting DMP.

6 Reasoning, Argument, and Proof

Reasoning is included in argument, but argument is a wider notion. Argument is reasoning used to try to resolve some central issue that is unsettled (Prakken and Sartor 2006). As stated above, in reasoning there is always a set of propositions called start points (premises) and a single endpoint (conclusion). In an argument, the conclusion is always the claim made by one party that is doubted or is open to doubt by the other party. The other party may be a single person or an audience composed of more than one person, for example a jury. In argument, the conclusion is always unsettled, or open to doubt. Indeed, that is the whole point of using an argument. If there is no doubt about a proposition, and everybody accepts it as true, there is no reason for arguing either for or against it. Thus, the speech act of putting forward an argument is different from the speech act of putting forward an explanation. An explanation is only appropriate if the proposition to be explained is taken as an accepted fact by all parties. An argument is only appropriate if the proposition at issue is not taken as an accepted fact by all parties, so that there is some doubt about whether it is true or not. The definition of the notion of an argument put forward here is dialectical, implying that an argument is only appropriate in a setting where two or more parties take part in trying to use reasoning to examine and evaluate the evidence on both sides of a disputed issue. The term "argumentation" is appropriately used here, because an argument, defined in this way, always needs to be evaluated within a procedural setting. There needs to be an opening stage, in which the ultimate issue needs to be specified, an argumentation stage where the arguments on both sides are

Table 1 Seven basic types of dialogue

Dialogue type	Initial situation	Participant's goal	Communal goal
Persuasion	Conflict of opinions	Persuade other party	Resolve issue
Inquiry	Need to have proof	Verify evidence	Prove hypothesis
Discovery	Need an explanation	Find a hypothesis	Support hypothesis
Negotiation	Conflict of interests	Get what you want	Settle issue
Information	Need information	Acquire information	Exchange information
Deliberation	Practical choice	Fit goals and actions	Decide what to do
Eristic	Personal conflict	Hit out at opponent	Reveal deep conflict

put forward so that they can be critically questioned by the opposing side, and the closing stage where it can be determined which side had the stronger argument.

One of the best examples of argumentation that can be given is a common law trial in which one side has made a claim that the other side disputes. The claim made by the first part is the ultimate *probandum*, the ultimate conclusion to be proved, while the other party has the job of casting doubt on the first party's argument. In this setting, there are three parties involved, the two primary parties and a third-party trier, who may be a judge or jury. Procedural rules determine what sort of evidence is admissible, and the two primary parties are supposed to use this evidence to try to prove their own contentions and cast doubt on the contentions of the other side. The trier makes a decision at the closing stage by examining the arguments on both sides and determining whether one side or the other successfully carried out its central task. The overall framework of a common law trial is that of a persuasion dialogue, because the party who has made the initial claim that defines the issue to be settled has a burden of persuasion. It has the burden to prove its claim by sufficient evidence or it loses by default and the other party wins. This burden is often called the presumption of innocence in criminal trials.

Although persuasion dialogue is a very important type of dialogue, so much so that argumentation is often mainly associated with it, there are other types of dialogue that are important from a legal reasoning point of view. Six types were distinguished in Walton and Krabbe (1995, 66), and the seventh type was also later recognized. Each type has a communal goal that needs to be distinguished from the individual goals of the participants. The defining characteristics of each type are shown in Table 1.

Each type of dialogue has an opening stage, an argumentation stage, and a closing stage. A dialogue is formally defined as an ordered 3-tuple $\{O, A, C\}$ where O is the opening stage, A is the argumentation stage, and C is the closing stage (Gordon and Walton 2009, 5). Procedural rules define what types of moves are allowed by the parties during the argumentation stage (Walton and Krabbe 1995). At the opening stage, the participants agree to take part in some type of dialogue that has an identified collective goal. The initial situation is framed at the opening stage, and the dialogue moves through the argumentation stage toward the closing stage. This way of abstractly modeling argumentation in a dialogue setting is normative, meaning

that it prescribes how the dialogue should ideally go if the participants aim to use reasoning to resolve the issue that was framed at the opening stage. In any real case, the actual order of events may be quite different.

In these dialogue systems, the simplest cases are those with only two participants. To test out the systems, Lodder (1999, 99) considered a short sample dialogue of legal argumentation arising out of an actual case. In this case, a gang member named Tyrell was searched during a football game, and he was found to be in possession of illegal drugs. The issue was whether this evidence had been obtained illegally. The two participants in the dialogue are called Bert and Ernie, but any names could be chosen, like Proponent and Respondent, or Black and White.

Bert: It was not allowed to search Tyrell.

Ernie: Why do you think so?

Bert: Only if someone is a suspect may he be searched, and Tyrell was not a suspect.

Ernie: I agree, but Tyrell was on probation, and had to allow a search at any time.

Bert: You are right, a search was allowed.

What is especially interesting about the argumentation in this dialogue is that it is based on defeasible reasoning, based on generally accepted legal rules that are subject to exceptions. Two of these general rules can be identified as follows.

Search Rule: Someone may be searched only if he is a suspect.

Probation Rule: Someone who is on probation can be searched at any time.

The reason behind the second rule is that one of the conditions of probation is to allow a search at any time. Turning to the dialogue, the line of argumentation can be followed along, based on how these defeasible rules are applied. First, Bert made the claim that it was not allowed to search Tyrell. This assertion can be taken to mean that Bert has advocated a viewpoint, to use the language of the critical discussion. In other words, he has made a statement and committed himself to the truth or acceptability of it. In the critical discussion model, it follows that Bert is obliged to defend his claim, if challenged, or give it up, if the other party can give a convincing argument against it. Ernie then questions Bert's claim, at the next move in the dialogue. Bert then fulfills his obligation by presenting an argument. His argument cites the claimed fact that Tyrell was not a suspect, along with the search rule as an additional premise. The two premises function together as an argument to support argument against Ernie's prior argument. Also, it should be noted that the search rule is a defeasible rule that is subject to exceptions. Thus Bert's argument, although it appears to be reasonable as matters stand, could be defeated by new evidence that might come into the dialogue. And in fact, that is what happens in the dialogue, when Ernie cites the presumed fact that Tyrell was on probation, and uses it along with the probation rule to put forward a new defeasible argument that defeats Bert's prior argument.

Information-seeking dialogue is, at first, hard to grasp as a framework of legal argumentation. The most familiar framework is that of persuasion dialogue, which is highly adversarial, in that its basic structure is based on opposed arguments. Prakken (2006) has investigated a formal framework for the study of dialogue games for argumentation in which each dialogue move either attacks or surrenders to some earlier

move of the other participant. This framework assumes an explicit reply structure on dialogues and imposes strict protocols on turn taking and relevance. However, Prakken has also explored less rigorous and more permissive formal dialogues in which these strong conditions are relaxed. In these more permissive types of dialogue, alternative replies to the same turn are allowed, and some dialogues do not have to adhere to the rule that each move must have a bearing on the outcome of the dialogue. Prakken (2006, 28) presented the following example dialogue.

Witness: Suspect was at home with me that day.

Prosecutor: Are you a student?

Witness: Yes.

Prosecutor: Was that day during summer holiday?

Witness: Yes.

Prosecutor: Aren't all students away during summer holiday?

Prakken cites this example as a case where the cross-examination of a witness has the goal of revealing an inconsistency in the testimony. He offered the example as a typical case in which a line of questioning does not indicate from the start where it is aiming. But clearly it is as species of information-seeking dialogue. It can also be identified as falling under the category of examination dialogue, a subspecies of information-seeking dialogue. It should probably be seen as taking place at the beginning stage of the line of questioning, in which the prosecutor is asking questions of a kind that attempt to pin down the commitments of the witness in such a way that they can be subsequently be critically examined. One such strategy of cross-examination that the prosecutor may have in mind is that of later trapping the witness into committing himself to an inconsistency, or to some proposition that would appear to be implausible to the judge or jury in the trial. On this model even though the dialogue is a fragmentary one, and very little context is given, it can be straightforwardly categorized as an instance of a kind of examination dialogue that falls under the information-seeking category. As seen from a viewpoint of persuasion dialogue, it may appear to be aimless, but as seen from a viewpoint of examination dialogue, it can be analyzed as part of a dialogue that has a goal.

So far there has been a lot of discussion about both reasoning and argument, but now we also need to consider the concept of proof. A proof may be defined as an argument in which the premises furnish sufficient evidence to reasonably accept the conclusion. A burden of proof is a requirement set on one side or the other to meet a standard of proof in order for the argument of that side to be judged successful as a proof (Gordon et al. 2007). The problem in legal reasoning, and indeed in all kinds of reasoning generally, is that it is rarely if ever possible to be able to prove a conclusion beyond all doubt. Hence, the question of when a burden of proof is met by a sequence of argumentation in a given case depends on the proof standard that is required for a successful argument in that case. What proof standard is required depends on the type of dialogue. In inquiry dialogue, for example an investigation into the cause of an air disaster has a very high standard of proof. In another dialogue

setting, the standard may not need to be set as high. Proof standards need to be set at the opening stage of a dialogue.

According to the scintilla of evidence standard, an argument is taken to be a proof even if there is only a small amount of evidence supporting it (Gordon and Walton 2009). The preponderance of evidence proof standard is met by an argument that is stronger than the opposed argument in the case, even if it is only slightly stronger. The clear and convincing evidence standard is higher than that of the preponderance standard, but not as high as the highest standard, called proof beyond reasonable doubt. The beyond reasonable doubt standard is the strongest one, and it is applicable in criminal cases. The beyond reasonable doubt standard is often equated with the presumption of innocence in criminal cases (Tillers and Gottfried 2006).

Burden of proof is a slippery and vague notion in law, and so far it has resisted any precise definition that has been unanimously accepted in law (Prakken and Sartor 2009). However, recent work in artificial intelligence has constructed logical models based on defeasible logic that are helpful in clarifying how the notion of burden of proof works in legal argumentation by distinguishing different kinds of burden of proof that can appear at different stages in a dialogue (Prakken and Sartor 2003). At the opening stage, when a person makes a claim at the first point in the sequence described above, he has a right to a legal remedy if he can bring forward facts that are sufficient to prove that he is entitled to some remedy. This is called the burden of claiming. The second type of onus is the burden of questioning. If one party makes an allegation by claiming that some proposition is true during the process of the argumentation, and the other party fails to present a counterargument, or even to deny the claim, then that claim is taken to be implicitly conceded. This type of burden of proof is called the burden of questioning because it puts an obligation on the other party to question or contest a claim made by the other side, by asking the other side to produce arguments to support its claim. The third burden is called the burden of production or the burden of producing evidence. It is the burden to respond to a questioning of one's claim by producing evidence to support it (Prakken and Sartor 2009).

The fourth type of burden of proof is called the burden of persuasion in law. It is set by law at the opening stage of the trial and determines which side has won or lost the case at the end of the trial once all the arguments have been examined. The burden of persuasion works differently in a civil proceeding than in a criminal one. In a civil proceeding, the plaintiff has the burden of persuasion for all the claims he has made as factual, while the defendant has the burden for any exceptions that he has pleaded. In criminal law, the prosecution has the burden of persuasion for all facts of the case. These include not only the elements of the alleged crime, but also the burden of disproving defenses. For example, let us say that in a murder case in a particular jurisdiction, the prosecution has to prove that there was a killing and that it was done with malice aforethought. If the defendant pleads self-defense, the prosecution has to prove that there was no self-defense (Prakken and Sartor 2009).

The fifth type of burden is called the tactical burden of proof. It applies during the argumentation stage of the trial, when a lawyer pleading a case has to make strategic decisions on whether it is better to present an argument or not. This is a hypothetical

assessment made only by the advocates on the two sides, and it can shift back and forth during a sequence of argumentation.

7 Conclusions

Some important general theoretical questions for argumentation theory underlie this chapter. Although it is generally appropriate to call the inferential structures studied in the chapter argumentation schemes, in many instances they could equally appropriately be called reasoning schemes. Discussion of this question in the chapter has led to another high-level question for argumentation theory: What is the difference between reasoning and argument? It appears that at least in some instances, argumentation schemes apply to cases as forms of inference used to generate conclusions by inference from a set of assumptions. In such cases, the conclusion may be drawn by the inference of a solitary agent and may not be disputed or doubted by a second party. In such cases, it may seem appropriate to better describe the argumentation scheme by calling it a reasoning scheme. So why are they called argumentation schemes at all, if the process of generating conclusions by inference would seem to be appropriately described by using the term "reasoning"? The answer given in this chapter is that they are forms of reasoning, and as such they can be used to identify instances of known kinds of reasoning, a useful task, but there is also an important reason why it is more generally appropriate to call them argumentation schemes. The reason is that the arguments fitting them are evaluated using sets of critical questions matching a particular scheme. This is a process in which one party, the proponent, puts forward an argument and another party, the respondent, asks critical questions in a dialogue format. In the simplest case, where two parties are involved, one of them is casting doubt on the other's argument. Here, the use of the term "argument" is highly appropriate, because for there to be an argument there has to be a claim that is unsettled.

References

Ashley, K. 1988. Arguing by analogy in law: A case-based model. In *Analogical reasoning*, ed. D.H. Helman, 205–224. Dordrecht: Kluwer.

Ashley, K. 2006. Case-based reasoning. In *Information technology and lawyers,* ed. A.R. Lodder and A. Oskamp, 23–60. Berlin: Springer.

Ashley, K. 2009. Ontological requirements for analogical, teleological and hypothetical reasoning. In *Proceeding of ICAIL 2009: 12th international conference on artificial intelligence and law*, 1–10. New York, N.Y.: Association for Computing Machinery.

Ashley, K., and E. Rissland. 2003. Law, learning and representation. *Artificial Intelligence* 150: 17–58.

Atkinson, K., and T. Bench-Capon. 2007. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence* 171: 855–874.

Atkinson, K., T. Bench-Capon, and P. McBurney. 2005. Arguing about cases as practical reasoning. In *Proceedings of the 10th international conference on artificial intelligence and law,* ed. G. Sartor, 35–44. New York, N.Y.: ACM Press.

Atkinson, K., T. Bench-Capon, and P. McBurney. 2006. Computational representation of practical argument. *Synthese* 152: 157–206.

Bench-Capon, T. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13: 429–448.

Bench-Capon, T. 2009. Dimension based representation of Popov v Hayashi. In *Modelling legal cases,* ed. K. Atkinson, 41–52. Barcelona: Huygens Editorial.

Bench-Capon, T. 2012. Representing Popov vs. Hayashi with Dimensions and Factors. *Artificial Intelligence and Law* 20: 15–35.

Bex, F. 2009. Analysing stories using schemes. In *Legal evidence and proof: statistics, stories, logic*, ed. H. Kaptein, H. Prakken and B. Verheij, 93–116. Farnham: Ashgate.

Brewer, S. 1996. Exemplary reasoning: Semantics, pragmatics and the rational force of legal argument by analogy. *Harvard Law Review* 923–1038.

Copi, I.M., and C. Cohen. 1998. *Introduction to logic*, 10th ed. Upper Saddle River: Prentice Hall.

Gordon, T.F. 2010. The Carneades argumentation support system. In *Dialectics, dialogue and argumentation*, ed. C. Reed, and C.W. Tindale. London: College Publications.

Gordon, T.F., and D. Walton. 2009. Proof burdens and standards. In *Argumentation and Artificial Intelligence*, ed. I. Rahwan, and G. Simari, 239–260. Berlin: Springer.

Gordon, T.F., H. Prakken, and D. Walton. 2007. The Carneades model of argument and burden of proof. *Artificial Intelligence* 171: 875–896.

Gray, B.E. 2002. Reported and recommendations on the law of capture and first possession: Popov v. Hayashi. In *Superior of the State of California for the City and County of San Francisco.* Case no. 400545, November 6, 2002. http://web.mac.com/graybe/Site/Writings_files/Hayashi%20Brief.pdf. Accessed 24 May 2009.

Guarini, M., A. Butchart, P. Simard Smith, and A. Moldovan. 2009. Resources for research on analogy: A multi-disciplinary guide. *Informal Logic* 29: 84–197.

Hamblin, C. 1970. *Fallacies*. London: Methuen.

Hart, H.L.A. 1949, 1951. The ascription of responsibility and rights. *Proceedings of the Aristotelian Society* 49: 171–194. Reprinted in *Logic and language,* ed. A. Flew. 145–166. Oxford: Blackwell, 1951.

Hart, H.L.A. 1961. *The concept of law*. Oxford: Oxford University Press.

Josephson, J.R., and S.G. Josephson. 1994. *Abductive inference: Computation, philosophy, technology*. New York, N.Y.: Cambridge University Press.

Levi, E.H. 1949. *An introduction to legal reasoning*. Chicago, Ill: University of Chicago Press.

Lodder, A.R. 1999. *Dialaw: On legal justification and dialogical models of argumentation*. Dordrecht: Kluwer.

Macagno, F., and D. Walton. 2009. Argument from analogy in law, the classical tradition, and recent theories. *Philosophy & Rhetoric* 42: 154–182.

McCarthy, K.M. 2002. Statement of decision. Case of Popov v. Hayashi #4005545. *Superior Court of California.* www.findlaw. Accessed 12 Dec 2002.

Nute, D. 1994. Defeasible logic. In *Handbook of logic in artificial intelligence and logic programming*, vol. 3. *Nonmonotonic reasoning and uncertain reasoning*, ed. D.M. Gabbay, C.J. Hogger, and J.A. Robinson, 353–395. Oxford: Oxford University Press.

Pardo, M.S., and R.J. Allen. 2008. Juridical proof and the best explanation. *Law and Philosophy* 27: 223–268.

Pennington, N., and R. Hastie. 1993. The story model for juror decision making. In *Inside the juror: The psychology of juror decision making*, ed. R. Hastie, 192–221. Cambridge: Cambridge University Press.

Pollock, J. 1995. *Cognitive Carpentry*. Cambridge, Mass.: MIT Press.

Prakken, H. 2005. AI & law, logic and argument schemes. *Argumentation* 19: 303–320.

Prakken, H. 2006. Models of persuasion dialogue. http://www.cs.uu.nl/groups/IS/archive/henry/argbookhp.pdf. (Originally published as Formal systems for persuasion dialogue. *The Knowledge Engineering Review* 21: 163–188, 2006.).

Prakken, H., and G. Sartor. 2006. A dialectical model of assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law* 4: 331–368.

Prakken, H., and G. Sartor. 2009. A logical analysis of burdens of proof. In *Legal evidence and proof: Statistics, stories, logic*, ed. H. Kaptein, H. Prakken, and B. Verheij, 223–253. Farnham: Ashgate Publishing.

Reed, C., and D. Walton. 2003. Diagramming, argumentation schemes and critical questions. In *Anyone who has a view: Theoretical contributions to the study of argumentation*, ed. F.H. van Eemeren, et al., 195–211. Dordrecht: Kluwer.

Sartor, G. 2005. *Legal reasoning: A cognitive approach to the law*. Berlin: Springer.

Schauer, F. 1987. Precedent. *Stanford Law Review* 39: 571–605.

Schauer, F. 2009. *Thinking like a lawyer*. Cambridge, Mass.: Harvard University Press.

Tillers, P. 1989. Webs of things in the mind: A new science of evidence. *Michigan Law Review* 87: 1225–1258.

Tillers, P., and J. Gottfried. 2006. Case comment—United States v. Copeland, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A collateral attack on the legal maxim that proof beyond a reasonable doubt is unquantifiable? *Law, Probability and Risk* 5: 135–157.

Twining, W., and D. Miers. 2010. How to do things with rules. Cambridge: Cambridge University Press.

Verheij, B. 2001. Legal decision making as dialectical theory construction with argumentation schemes. In *The 8th international conference on artificial intelligence and law: Proceedings of the conference,* 225−236. New York Association for Computing Machinery. http://www.ai.rug.nl/~verheij/publications.htm.

Walton, D. 1990a. What is reasoning? What is an argument? *Journal of Philosophy* 87: 399–419.

Walton, D. 1990b. *Practical reasoning*. Savage, Md.: Roman and Littlefield.

Walton, D. 1997. *Appeal to expert opinion*. University Park, Penn.: Penn State Press.

Walton, D. 2002. Are some modus ponens arguments deductively invalid? *Informal Logic* 22: 19–46.

Walton, D. 2008. *Witness testimony evidence: Argumentation, artificial intelligence and law*. Cambridge: Cambridge University Press.

Walton, D. 2010. Similarity, precedent and argument from analogy. *Artificial Intelligence and Law* 18: 217–246.

Walton, D., and E.C.W. Krabbe. 1995. *Commitment in dialogue*. Albany, Texas: SUNY Press.

Walton, D., and T.F. Gordon. 2005. Critical questions in computational models of legal argument. In *IAAIL workshop series, international workshop on argumentation in artificial intelligence and law*, ed. P.E. Dunne, and T.J.M. Bench-Capon, 103–111. Nijmegen: Wolf Legal Publishers.

Walton, D., C. Reed, and F. Macagno. 2008. *Argumentation schemes*. Cambridge: Cambridge University Press.

Weinreb, L.L. 2005. *Legal reason: The use of analogy in legal argument*. Cambridge: Cambridge University Press.

Wigmore, J.H. 1931. *The principles of judicial proof*, 2nd ed. Boston, Mass.: Little, Brown and Company.

Wigmore, J.H. 1940. *Evidence in trials at common law*. Boston, Mass.: Little, Brown & Co.

Wooldridge, M., and W. van der Hoek. 2005. On obligations and normative ability: Towards a logical analysis of the social contract. *Journal of Applied Logic* 3: 396–420.

Wyner, A., and T. Bench-Capon. 2007. Argument schemes for legal case-based reasoning. In *Legal knowledge and information systems (JURIX 2007)*, ed. A. Lodder, and L. Mommers, 139–149. Amsterdam: IOS Press.

A. Wyner, T.J.M. Bench-Capon, and K. Atkinson. 2007. Arguments, values and baseballs: Representation of Popov v. Hayashi. In *Legal knowledge and information systems (JURIX 2007)*, eds. A. Lodder and L. Mommers, 151−160. Amsterdam: IOS Press.

Norms in Action: A Logical Perspective

Emiliano Lorini

1 Introduction

A theory of action is fundamental for legal theory, as the law is meant to direct behaviour: it influences the behaviour of agents who can understand the law's prescriptions and act accordingly. A connection between law and action is assumed by the most diverse approaches to the law; when no reference is made to this connection it is since it appears to be an obvious truism. Let us list just a few examples where this connection appears most clearly.

The Digest of Justinian, while not explicitly linking law to action, distinguishes the way in which the law influences human action, by affirming that the law is meant to "command, forbid and punish" (D 1.7). Similarly, Cicero affirmed that the (just or rational) law "enjoins what ought to be done and forbids the opposite," i.e., that it has the function to "enjoin the right action and forbid wrong-doing" (Cicero 1998, n. 18).

The connection between law and action is also explicitly included in Aquinas's definition of the law as "a certain rule and measure of acts whereby man is induced to act or is restrained from acting" (Aquinas 1947, q90, a1).

Similarly, Grotius affirmed that "the law of nature is a dictate of right reason, which points out that an act, according as it is or is not in conformity with rational nature, has in it a quality of moral baseness or moral necessity; and that, in consequence, such an act is either forbidden or enjoined" (Grotius [1625] 1925, Book I, chap. I.x.1).

Leibniz also affirmed the connection between law and action (and between virtues and obligations of natural law) affirming that the legal obligation concerns "what it

The introduction to this chapter has been written by Emiliano Lorini and Giovanni Sartor.

E. Lorini (✉)
IRIT-CNRS Toulouse University, Toulouse, France
e-mail: Lorini@irit.fr

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_4

necessary for a good man to do," while permissibility covers "what it possible for a a good man to do" (Leibniz [1671] 1930, 431).

Moving from natural law to nineteenth-century positivism, we can find the connection between law and action in Jeremy Bentham's definition of law: "A law may be defined as an assemblage of signs declarative of a volition conceived or adopted by the sovereign in a state, concerning the conduct to be observed in a certain case by a certain person or class of persons" (Bentham [1872] 1970, 1).

The connection between law and action can also be found in Hans Kelsen's statement that "the norms of an order regulate […] always human behavior—only it can be regulated by norms" (Kelsen 1967, 14–15). According to Kelsen, "the behavior regulated by a normative order is either a definite action or the omission (nonperformance) of such an action." In Kelsen's perspective sanctions are connecting to actions in two ways: they are meant to induce the omission of the punished actions, and they are established by authorizing "coercive action" (i.e., the imposition of sanctions) by state authorities.

The connection between law and action, action being the matter of legal regulations, is also affirmed by H. L. A Hart. For this author, both primary and secondary norms are concerned with actions: primary duty-imposing norms are meant to direct human actions and, on the other hand, secondary power-conferring norms also provide for the institutional effects of the actions through the exercise of such powers.

> Under rules of the one type, which may well be considered the basic or primary type, human beings are required to do or abstain from certain actions, whether they wish to or not. Rules of the other types are in a sense parasitic upon or secondary to the first; for they provide that human beings may by doing or saying certain things introduce new rules of the primary type, extinguish or modify old ones, or in various ways determine their incidence or control their operations. Rules of the first type impose duties; rules of the second type confer powers, public or private (Hart 1994).

In legal theory, a necessary connection between law and action is also affirmed by those approaches that view legal norms, and in particular duty-imposing norms as "reasons for action," at least when such norms are issued by a legitimate authorities (Raz 1979).

Besides being discussed or assumed in legal theory, the connection between law and action, has been the object of a vast doctrinal discussion in different domains of the law. In private law, it is often addressed in connection with the distinction of different kinds of triggers for legal effects. In particular, merely natural facts are traditionally distinguished from human acts, the latter consisting of behaviour being controlled, or at least controllable, by the agent. Within human acts, declarations of will (also called juristic acts) are often distinguished: they are acts consisting in declarations of legal outcomes (obligations or transfers between the parties) which are meant to produce such outcomes. Contracts are the most notable class of the latter acts. Criminal law focuses on the structure of criminal action, which includes discussions of the connection between the behavioural components (the so-called actus reus) and the corresponding mental states or attitudes (mens rea).

While the rich legal tradition of legal theory and legal doctrine provide many ideas for formal analysis of action—which cannot be addressed in this contribution—very

few logical accounts of action have been so far provided by legal scholarships. Only a few contribution—at the interface of legal theory, philosophical logic, and, more recently, computing—have attempted at formalizing the action component of legal norms.

In particular, the concept of an act is at the core of the logical analysis of legal prescriptions developed by Von Wright (1963), the founder of modern deontic logic. According to this author, the law consists of behavioural prescriptions, i.e., of "commands, permissions, and prohibitions, which are given or issued to agents concerning their conduct." Correspondingly, legal norms are modelled by applying deontic modal operators for obligation and permission to action descriptions. Von Wrights's logic of action is connected to the idea of a transition between the states of affairs existing before and after the act: "acts may quite appropriately be described as the bringing about or effecting (at will?) of a change." Two types of acts are distinguished, actions and forbearances, which consist in bringing about, or in refraining from bringing about such changes.

The logical connection between laws and action was also addressed by Alchourrón and Bulygin (1971), who postulate that a set of possible actions is available (the universe of action), whose truth-functional compounds provide possible contents for deontic operators.

A critical analysis of Von Wright's theory was provided by Ota Weinberger (1998), a leading legal theorist and logician, who argued that the law requires a different approach to action, which he calls a "formal-finalistic action theory." According to this approach, an action is viewed as the outcome of a choice based on an information process, whose input includes both factual and normative information. This process involves the solution of an optimization problem, which may require taking into account multiple goals as well as constraints over means. Unfortunately, Weiberger did not provide a formal model of its theory of action.

A formally developed logic of norm and action, meant to address legal content, was developed by Kanger (1972). After characterizing "a system of law" as "a set of rules which has the purpose of regulating human action under certain conditions of law," Kanger modelled the norms of such a system by combining deontic logic and action logic. The basic idea of Kanger's approach to action is that an agent brings about a result—she does the action of bringing it about—when the agent's behaviour is both a necessary and a sufficient condition for the result to be produced. Action description of this kind can then be the object of deontic operators. Kanger's approach has been refined and developed in particular by Pörn (1977), who defined a logic of "bringing it about that" (BIAT), and developed a modal semantics for it. According to this logic, agent i brings about a state of affairs φ if and only if two conditions hold: it is necessary for everything that i does that φ, and but for what i does it might be the case that $\neg\varphi$. In other terms given i's behaviour necessarily φ is the case, and without that behaviour $\neg\varphi$ might be the case. For instance, it may be said that I close the door if, given my behaviour, necessarily the door is closed and, without my behaviour, it might be open. Further developments of the BIAT logic have been proposed by a number of authors, among which Elgesem (1997), see also Governatori and Rotolo (2005). A very rich traditions, which we cannot examine

in this paper concerns the combination of the BIAT action logic and Hohfeldian modalities, on which see (Lindahl 1977; Jones and Sergot 1996).

Only recently logics of action have been provided that are sufficiently general to represent norm-related concepts such as causality, responsibility, and influence. Such logics are needed to capture aspects of legal agency that so far have only been addressed by informal doctrinal accounts. In fact, to describe social interaction in a formal way, it is necessary to have a representation language that allows to describe, at the same time, the causal relations between the actions and their effects, the agents' action repertoires and capabilities as well as the effects of the joint actions of agents. Actions occur in time and have a duration. Thus, a logical theory of interaction requires a clear understanding of the relationship between actions and time.

One of the logics that has been used to represent the concepts of individual and joint actions is propositional dynamic logic (PDL) (Harel and Tiuryn 2000). PDL has been introduced in theoretical computer science about thirty years ago in order to represent the concept of a (computer) program and the basic operations on programs (e.g., sequential composition, non-deterministic choice, iteration, test). Consequently, the use of PDL in modelling interaction between agents (see, e.g., Schmidt et al. 2004) works under the assumption that actions can be conceived as programs executed by the agents in the system. PDL's semantics is based on the concept of labelled transition system, that is to say, a graph whose vertices represent possible states of the system and whose edges are labelled with actions of agents. These edges represent transitions between states that are determined by the execution of an atomic program. PDL has been also shown to be a valuable formal language for representing normative concepts such as obligation, prohibition and permission (Meyer 1988; van der Meyden 1996). In PDL atomic programs are abstract entities in the sense that their semantics is just specified in terms of state transitions.

A concrete variant of PDL, called DL-PA (Dynamic Logic of Propositional Assignments) (Balbiani et al. 2013; Tiomkin and Makowsky 1985), has also been proposed. Differently from PDL, in DL-PA atomic programs are concrete. Specifically, they are assignments of propositional variables (i.e., an atomic program consists in setting to either $\top$ or $\bot$ the value of a given propositional variable p). The DL-PA notion of action, viewed as a propositional assignment, is compatible with Von Wright's idea of viewing an action as bringing about or effecting (at will) of a change. It is also shared with other formal systems proposed in the recent years in artificial intelligence (AI) such as boolean games and the Coalition Logic of Propositional Control (CL-PC). CL-PC was introduced by van der Hoek and Wooldridge (2005) as a formal language for reasoning about capabilities of agents and coalitions in multi-agent environments. In this logic, the notion of capability is modelled by means of the concept of *control*. In particular, it is assumed that each agent i is associated with a specific finite subset Atm_i of the finite set of all atomic propositions Atm. Atm_i is the set of propositions *controlled* by the agent i. That is, the agent i has the ability to assign a (truth) value to each proposition Atm_i but cannot affect the truth values of the propositions in $Atm \backslash Atm_i$. It is also assumed that control over propositions is exclusive, that is, two agents cannot control the same proposition (i.e., if $i \neq j$ then $Atm_i \cap Atm_j = \emptyset$). Moreover, it is assumed that control over

propositions is complete, that is, every proposition is controlled by at least one agent (i.e., for every $p \in Atm$ there exists an agent i such that $p \in Atm_i$).

Boolean games (Harrenstein et al. 2001; Bonzon et al. 2006) share with CL-PC the idea that an agent's action consists in affecting the truth values of the variables she controls. They are games in which each player wants to achieve a certain goal represented by a propositional formula: they correspond to the specific subclass of normal form games in which agents have binary preferences (i.e., payoffs can be either 0 or 1). They have been proved to provide a useful and natural abstraction for reasoning about social interaction in multi-agent systems.

An alternative approach to the logical formalization of action and of the connection between actions and norms in multi-agent domains is the logic STIT (the logic of "seeing to it that"), which was introduced for the first time in the philosophical area (Belnap et al. 2001; Horty 2001; Horty and Belnap 1995) and has become popular in AI in the recent years. This logic is well-suited to represent the concept of causality (whether an agent brings about a certain state of affairs as a result of her current choice) as well as social concepts such as the concepts of responsibility, guilt, delegation, and social influence that are of primary importance in modelling social relations between human and artificial agents. Two variants of STIT have been studied in the literature which differ at the syntactic level: an atemporal version and a temporal version. The temporal version of STIT is a combination of temporal operators of temporal logic for expressing temporal properties of facts (e.g., whether a given fact φ will be true in the future) and operators of agency that allow to express the consequences of the choice of an agent or group of agents. The language of atemporal STIT is nothing but the language of temporal STIT restricted to the agency operators which does not include temporal operators. A central assumption of STIT is that agents' choices are independent, in the sense that an agent can never be deprived of choices due to the choices made by other agents. This distinguishes STIT from the BIAT approach in which agents' choices are not necessarily independent (see Sergot 2014 for a discussion on this point).

Other logical systems have been proposed in the recent years which move from the concept of action to the game-theoretic concept of strategy. Informally speaking, a strategy for a certain agent specifies, for every state of the system characterized by a tree or by a transition system, what the agent is expected to do at this state of the system. The most representative example of strategy logics is alternating-time temporal logic (ATL) (Alur et al. 2002) which can be seen as the strategic variant of Coalition Logic (CL) by Pauly (2002) and which allows to formally represent the consequences of the strategy of a certain agent or coalition of agents.[1]

Plan of the chapter The aim of the present chapter is to show how logic can be used for formalizing legal concepts in which the notion of action plays a fundamental role. Given the diversity of existing logics of action, we have decided to focus on STIT logic. There are two main motivations behind this decision. First of all, STIT offers a general framework for modelling action and time and for representing the

[1] Differently from STIT, CL can only represent the consequences of the choice of a certain player or coalition players, a choice being the restriction of a strategy to the current state of the system.

consequence of an agent's choice. As we will show, the latter captures a fundamental aspect of agent causation that is relevant for legal theory. Secondly, the formal semantics of STIT is extremely elegant and intuitive. Moreover, it is directly connected with the formal representation of action and action-related concepts (e.g., power and capability) used in game theory. From this perspective, STIT has a high-level generality, as it can be seen as the prototypical logic of action based on a game-theoretic semantics.

The paper is organized as follows. Section 2 is devoted to illustrate in a rather informal way the semantics of STIT as well as its formal language. Section 3 is devoted to illustrate the use of STIT for the formalization of responsibility and influence, two concepts that are highly relevant for legal theory. Finally, Sect. 4 presents a simple extension of STIT by the concept of "obligation to do" based on a utilitarian view of norms.

2 A Logic for Reasoning About Choices, Actions, and Time

STIT logic (the logic of *seeing to it that*) by Belnap et al. (2001) is one of the most prominent formal accounts of agency. It is the logic of sentences of the form "the agent i sees to it that φ is true." Different semantics for STIT have been proposed in the literature (Belnap et al. 2001; Broersen 2011; Wölf 2002; Schwarzentruber 2012). The original semantics of STIT by Belnap et al. (2001) is defined in terms of **BT+AC** structures: branching time structures (**BT**) augmented by agent choice functions (**AC**). A **BT** structure is made of a set of moments and a tree-like ordering over them. An **AC** for an agent i is a function mapping each moment m into a partition of the set of histories passing through that moment, a history h being a maximal set of linearly ordered moments and the equivalence classes of the partition being the possible choices for agent i at moment m.

In Lorini (2013), a Kripke-style semantics for STIT has been proposed. On the conceptual side, the main difference between this Kripke semantics for STIT and Belnap et al.'s **BT+AC** semantics is that the former takes the concept of *world* as a primitive instead of the concept of *moment* and defines: (i) a *moment* as an equivalence class induced by a certain equivalence relation over the set of worlds, (ii) a *history* as a linearly ordered set of worlds induced by a certain partial order over the set of worlds, and (iii) an agent i's set of *choices* at a moment as a partition of that moment. The main advantage of the Kripke semantics for STIT over Belnap et al.'s original semantics in terms of **BT+AC** structures is that the former is a standard multi-relational semantics commonly used in the area of modal logic (Blackburn et al. 2001), whereas the latter is non-standard.

It is worth noting that, at the semantic level, temporal STIT can be conceived as a logic of action interpreted over infinitely repeated games. This highlights the connection between STIT and game theory.

The Kripke semantics of STIT is illustrated in Fig. 1, where each moment m_1, m_2, and m_3 consists of a set of worlds represented by points. For example, moment

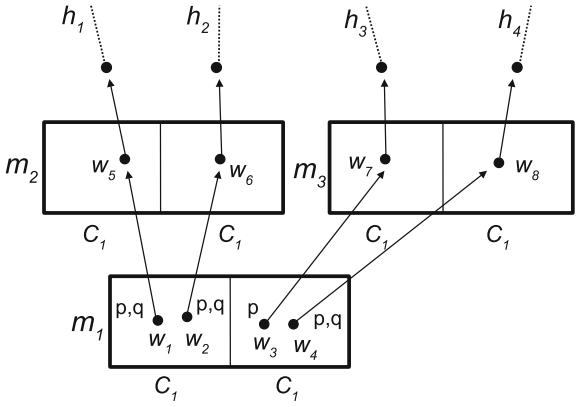


Fig. 1 Illustration of Kripke semantics of STIT

m_1 consists of the set of worlds $\{w_1, w_2, w_3, w_4\}$. Moreover, for every moment there exists a set of histories passing through it, where a history is defined as a linearly ordered set of worlds. For example, the set of histories passing through moment m_1 is $\{h_1, h_2, h_3, h_4\}$. Finally, for every moment, there exists a partition which characterizes the set of available choices of agent 1 in this moment. For example, at moment m_1, agent 1 has two choices available, namely $\{w_1, w_2\}$ and $\{w_3, w_4\}$. Note that an agent's set of choices at a certain moment can also be seen as a partition of the set of histories passing through this moment. For example, we can identify the choices available to agent 1's at m_1 with the two sets of histories $\{h_1, h_2\}$ and $\{h_3, h_4\}$. Following Horty (2001), the Kripke semantics for STIT also account for collective choices of groups of agents. Specifically, the choice of a group coincides with the intersection of the choices of the agents in the group. For instance, in Fig. 2, the individual choices of agents 1 and 2 are, respectively, $\{w_1, w_2, w_5, w_6\}$ and $\{w_1, w_2, w_3, w_4\}$, while the collective choice of group $\{1, 2\}$ is $\{w_1, w_2\}$.

Clearly, for every moment m in a Kripke semantics for STIT, one can identify the set of histories passing through it by considering all histories that contain at least one world in the moment m. Moreover, an agent i's set of choices available at m can also be seen as a partition of the set of histories passing through m. At first glance, an important difference between Belnap et al.'s semantics and Kripke semantics for STIT seems to be that in the former the truth of a formula is relative to a moment-history pair m/h, also called *index*, whereas in the latter it is relative to a world w. However, this difference is only apparent, because in the Kripke semantics for STIT there is a one-to-one correspondence between worlds and indexes, in the sense that: (i) for every index *m/h* there exists a unique world w at the intersection between m and h, (ii) and for every world w there exists a unique index *m/h* such that the intersection between m and h includes w.

In the Kripke semantics for STIT the concept of world should be understood as a "time point" and the equivalence class defining a moment as a set of alternative concomitant "time points." In this sense, the concept of moment captures a first aspect of indeterminism, as it represents the alternative ways the *present* could be.

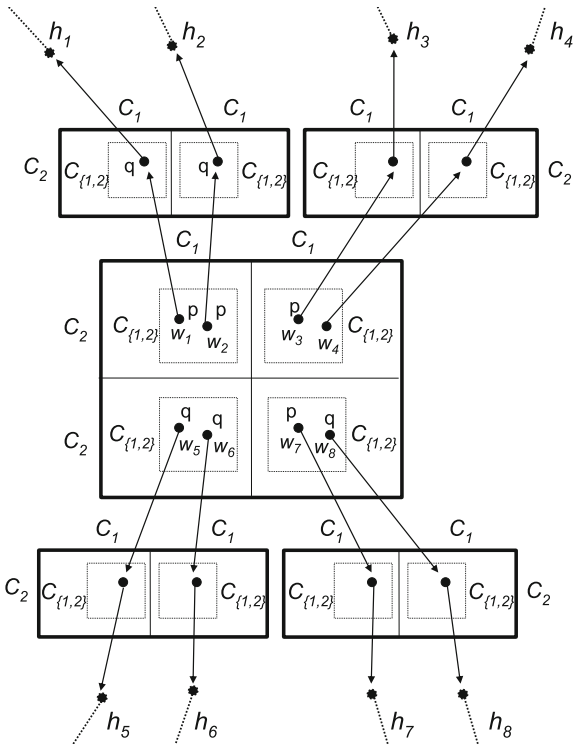


Fig. 2 Kripke semantics of STIT with groups

A second aspect of indeterminism is given by the fact that moments are related in a (tree-like) branching time structure. In this sense, the *future* could evolve in different ways from a given moment. In the Kripke semantics for STIT these two aspects of indeterminism are related, as illustrated in Fig. 1. Indeed, if two distinct moments m_2 and m_3 are in the future of moment m_1, then there are two distinct worlds in m_1 (w_1 and w_3) such that a successor of the former (w_5) is included in m_2 and a successor of the latter (w_7) is included in m_3.

In Horty (2001), the following language of temporal STIT (TSTIT), i.e., the variant of STIT with tense operators, is considered:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid [J \text{ stit}]\varphi \mid \Box\varphi \mid \mathsf{G}\varphi \mid \mathsf{H}\varphi$$

where p ranges over an infinite set of atomic propositions Atm and J ranges over a set of groups of agents $X \subseteq 2^{Agt} \setminus \{\emptyset\}$, where Agt is a finite set of agents whose elements are denoted by $i, j, \ldots$ For notational convenience, we write $[i \text{ stit}]$ instead of $[\{i\} \text{ stit}]$ with $i \in Agt$.

When $X = 2^{Agt} \setminus \{\emptyset\}$, the previous TSTIT language allows us to talk about time, the consequences of all agents' individual choices as well as the consequences of all groups of agents' collective choices.

Let us discuss the meaning of the different modal operators. The formal language of TSTIT includes the future tense operator G and the past tense operator H, where $\mathsf{G}\varphi$ and $\mathsf{H}\varphi$, respectively, stand for "φ will always be true in the future" and "φ has always been true in the past." For example, the formula $\mathsf{G}\neg p$ is true at world w_1 in Fig. 1. Indeed, it is the case that p is false at all future worlds of w_1. Moreover, the formula $\mathsf{H}p$ is true at world w_5 since it is the case that p is true at all past worlds of w_5. In Lorini and Sartor (2016), a variant of temporal STIT with discrete time is studied. It includes the "next" operator X (where $\mathsf{X}\varphi$ stands for "φ is going to be true in the next world") and the "yesterday" operator Y (where $\mathsf{Y}\varphi$ stands for "φ was true in the previous world").

Moreover, the previous TSTIT language also includes the so-called historical necessity operator $\Box$ which allows us to represent those facts that are necessarily true, in the sense of being true at every point of a given moment or, equivalently, at every history passing through a given moment. For example, the formula $\Box p$ is true at world w_1 in Fig. 1 since p is true at every point of moment m_1 including world w_1.

The "historical possibility" operator $\Diamond$ is defined to be the dual operator of $\Box$, $\langle J \text{ stit}\rangle$ is defined to be the dual operator of $[J \text{ stit}]$, while F and P are defined to be the dual operators of G and H. That is:

$$\begin{aligned}
\Diamond\varphi &\stackrel{\text{def}}{=} \neg\Box\neg\varphi\\
\langle J \text{ stit}\rangle\varphi &\stackrel{\text{def}}{=} \neg[J \text{ stit}]\neg\varphi\\
\mathsf{F}\varphi &\stackrel{\text{def}}{=} \neg\mathsf{G}\neg\varphi\\
\mathsf{P}\varphi &\stackrel{\text{def}}{=} \neg\mathsf{H}\neg\varphi
\end{aligned}$$

STIT logic provides for different concepts of agency, all characterized by the fact that an agent acts only if she sees to it that a certain state of affairs is the case. In the previous TSTIT language, the so-called Chellas STIT operator $[i \text{ stit}]$, named after its proponent (Chellas 1992), is taken as a primitive. According to the STIT semantics, an agent i Chellas-sees-to-it that φ, denoted by formula $[i \text{ stit}]\varphi$, at a certain world w if and only if, for every world v, if w and v belong to the same choice of agent i then φ is true at world v. For example, in Fig. 1, agent 1 Chellas-sees-to-it that p at world w_1 because p is true both at world w_1 and at world w_2. The previous TSTIT language generalizes the Chellas STIT operator to groups of agents. For example, suppose $Agt = \{1, 2\}$. Then, in Fig. 2, it is the case that group $\{1, 2\}$ Chellas-sees-to-it that p, denoted by formula $[\{1, 2\} \text{ stit}]p$, at world w_1, because $\{w_1, w_2\}$ corresponds to the collective choice of group $\{1, 2\}$ at w_1, and p is true both at world w_1 and at world w_2.

A more sophisticated operator of agency is the deliberative STIT (Horty and Belnap 1995) which is defined as follows by means of the Chellas STIT operator and the historical necessity operator $\Box$:

$$[i \text{ dstit}]\varphi \stackrel{\text{def}}{=} [i \text{ stit}]\varphi \wedge \Diamond\neg\varphi$$

In other words, deliberative STIT satisfies the same positive condition as Chellas STIT *plus* a negative condition: an agent i deliberately sees to it that φ, denoted by formula $[i \text{ dstit}]\varphi$, at a certain world w if and only if: (i) agent i Chellas-sees-to-it that φ at w, that is to say, agent i's current choice at w ensures φ, and (ii) at w agent i could make a choice that does not necessarily ensure φ.[2] Notice that the latter is equivalent to say that there exists a world v such that w and v belong to the same moment and φ is false at v. For example, in Fig. 1, agent 1 deliberately sees to it that q at world w_1 because q is true both at world w_1 and at world w_2, while being false at world w_3. In other terms, while the truth of $[i \text{ stit}]\varphi$ only requires that i's choice ensures that φ, the truth of $[i \text{ dstit}]\varphi$ also requires that i had the opportunity of making an alternative choice that would not guarantee that φ would be the case. Deliberative STIT captures a fundamental aspect of the concept of action, namely the idea that for a state of affairs to be the consequence of an action (or for an action to be the cause of a state of affairs), it is not sufficient that the action is a sufficient condition for that state of affairs to hold, it is also required that, without the action, the state of affairs possibly would not hold. In this sense, while $[i \text{ stit}]\varphi$ at w is consistent with (and is indeed entailed by) the necessity of φ at w, $[i \text{ dstit}]\varphi$ at w is incompatible with the necessity of φ at w, since it requires that at w also $\neg\varphi$ was an open possibility. Consequently, the deliberative STIT is more appropriate than the Chellas STIT to describe the consequences of an agent's action, as *incompatibility with necessity* is a requirement for any reasonable concept of action.[3]

In Lorini (2013), a sound and complete axiomatization for the fragment of the previous TSTIT language in which $X = \{\{i\} \mid i \in Agt\} \cup \{Agt\}$, with respect to the STIT Kripke semantics, is provided. It is summarized in Fig. 3.

This includes all tautologies of classical propositional calculus **(PC)** as well as modus ponens **(MP)**. Moreover, we have all principles of the normal modal logic S5 for every operator $[i \text{ stit}]$, for the operator $[Agt \text{ stit}]$ and for the operator $\Box$, all principles of the normal modal logic KD4 for the future tense operator G and all principles of the normal modal logic K for the past tense operator H. That is, we have Axiom K for each operator: $(\blacksquare\varphi \wedge \blacksquare(\varphi \rightarrow \psi)) \rightarrow \blacksquare\psi$ with $\blacksquare \in \{\Box, \mathsf{G}, \mathsf{H}, [Agt \text{ stit}]\} \cup \{[i \text{ stit}] \mid i \in Agt\}$. We have Axiom D for the future tense modality G: $\neg(\mathsf{G}\varphi \wedge \mathsf{G}\neg\varphi)$. We have Axiom 4 for $\Box$, G, $[Agt \text{ stit}]$ and for every $[i \text{ stit}]$: $\blacksquare\varphi \rightarrow \blacksquare\blacksquare\varphi$

[2] We shall not consider here the achievement STIT operator by Belnap and Perloff (1988) which provide a more sophisticated account of agency but whose interpretation is considerably more complicated than the semantics of the deliberative STIT.

[3] The classical argument against the use of Chellas STIT for modelling action is that, according to Chellas STIT, an agent brings about all tautologies and that it is counterintuitive to say that a tautology is a consequence of an agent's action.

PC	All tautologies of classical propositional calculus
S5(i)	All S5-principles for the operators $[i\ \mathsf{stit}]$
S5$(\Box)$	All S5-principles for the operator $\Box$
S5(Agt)	All S5-principles for the operator $[Agt\ \mathsf{stit}]$
KD4(G)	All KD4-principles for the operator G
K(H)	All K-principles for the operator H
$(\Box \to i)$	$\Box\varphi \to [i\ \mathsf{stit}]\varphi$
$(i \to Agt)$	$([1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge [n\ \mathsf{stit}]\varphi_n) \to [Agt\ \mathsf{stit}](\varphi_1 \wedge \ldots \wedge \varphi_n)$
(AIA)	$(\Diamond[1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge \Diamond[n\ \mathsf{stit}]\varphi_n) \to \Diamond([1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge [n\ \mathsf{stit}]\varphi_n)$
(Conv$_{\mathsf{G,H}}$**)**	$\varphi \to \mathsf{GP}\varphi$
(Conv$_{\mathsf{H,G}}$**)**	$\varphi \to \mathsf{HF}\varphi$
(Connected$_{\mathsf{G}}$**)**	$\mathsf{PF}\varphi \to (\mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi)$
(Connected$_{\mathsf{H}}$**)**	$\mathsf{FP}\varphi \to (\mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi)$
(NCUH)	$[Agt\ \mathsf{stit}]\mathsf{G}\varphi \to \mathsf{G}\Box\varphi$
(MP)	$\dfrac{\varphi,\ \varphi \to \psi}{\psi}$
(IRR)	$\dfrac{(\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)) \to \varphi}{\varphi}$, provided p does not occur in φ

Fig. 3 Axiomatization of for the TSTIT with agency operators $[i\ \mathsf{stit}]$ and $[Agt\ \mathsf{stit}]$

with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}], \mathsf{G}\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. Furthermore, we have Axiom T for $\Box$, $[Agt\ \mathsf{stit}]$ and for every $[i\ \mathsf{stit}]$: $\blacksquare\varphi \to \varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}]\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. We have Axiom B for $\Box$, $[Agt\ \mathsf{stit}]$ and for every $[i\ \mathsf{stit}]$: $\varphi \to \blacksquare\neg\blacksquare\neg\varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}]\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. Finally we have the rule of necessitation for each modal operator: $\frac{\varphi}{\blacksquare\varphi}$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}], \mathsf{G}, \mathsf{H}\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$.

$(\Box \to i)$ and **(AIA)** are the two central principles in Xu's axiomatization of the Chellas's STIT operators $[i\ \mathsf{stit}]$ (Xu 1998). According to Axiom $(\Box \to i)$, if φ is true regardless of what every agent does, then every agent sees to it that φ. In other words, an agent brings about those facts that are inevitable.[4] According to Axiom $(i \to Agt)$, all agents bring about together what each of them brings about individually.

We have principles for the tense operators and for the relationship between time and action. **(Connected**$_{\mathsf{G}}$**)** and **(Connected**$_{\mathsf{H}}$**)** are the basic axioms for the linearity of the future and for the linearity of the past (Goldblatt 1992). **(Conv**$_{\mathsf{G,H}}$**)** and **(Conv**$_{\mathsf{H,G}}$**)** are the basic interaction axioms between future and past of minimal tense logic according to which "what is, will always have been" and "what is, has always been going to be."

[4] Xu considers a family of axiom schemas **(AIA**$_k$**)** for independence of agents of the form $(\Diamond[1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge \Diamond[k\ \mathsf{stit}]\varphi_k) \to \Diamond([1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge [k\ \mathsf{stit}]\varphi_k)$ that is parameterized by the integer k. As pointed out by (Belnap et al. 2001), **(AIA**$_{k+1}$**)** implies **(AIA**$_k$**)**. Therefore, as Agt is finite, in OPDL the family of axiom schemas can be replaced by the single axiom **(AIA)**.

Axiom **(NCUH)** corresponds to so-called property of "no choice between undivided histories" which is implicit in the Kripke semantics for STIT illustrated above: if in some future world φ will be possible then the actual collective choice of all agents will possibly result in a state in which φ is true.

(IRR) is a variant of the well-known Gabbay's irreflexivity rule that has been widely used in the past for proving completeness results for different kinds of temporal logic in which time is supposed to be irreflexive (see, e.g., Gabbay et al. 1994; Zanardo 1996; Reynolds 2003; von Kutschera 1997). The idea is that the irreflexivity for time, although not definable in terms of an axiom, can be characterized in an alternative sense by means of the rule **(IRR)**. This rule is perhaps more comprehensible if we consider its contrapositive: if p does not occur in φ and φ is TSTIT consistent, then $\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p) \wedge \varphi$ is TSTIT consistent.

3 Formalization of Responsibility and Influence

This section is devoted to illustrate the application of STIT to the logical formalization of the concepts of responsibility (Sect. 3.1) and social influence (Sect. 3.2) which requires a comprehensive theory of the relationship between action and time.

3.1 Responsibility

The concept of responsibility is highly relevant not only for the legal domain but also for AI. Specifically, it has been proved to be useful in the domain of autonomous agents and multi-agent systems (MASs). For instance, autonomous vehicles should be endowed with the capability of reasoning about their own responsibility and that of others. This kind of capability allows agents to identify those actions that might be blameworthy, because they do not conform to legal norms, and therefore refrain from performing them. Moreover, an intelligent virtual agent interacting with a human can be designed to recognize humans' emotions such as guilt or pride and to act consequently. This specific capacity can be achieved by endowing the agent with the more general capability of reasoning about humans' responsibility and humans' beliefs about her own and others' responsibility.

In Lorini et al. (2014) a formalization of the concept of responsibility in STIT is proposed. This focuses on both the consequences that the agent's actual choices have for other agents and the consequences of the actions that the agent could have chosen to perform and did not. More generally, it considers both the active (the agent's seeing to it that φ is the case) and passive (the agent's preventing φ from happening) dimensions of responsibility. The former is also called *responsibility for action*, while the latter is also called *responsibility for omission*. In particular:

An agent i is actively responsible for ensuring that a certain fact is true if and only if, i sees to it that the fact is true, regardless of what the others have decided to do.

An agent i is passively responsible for ensuring that a certain fact is true if and only if, the fact is actually true and i could have prevented it from being true, regardless of what the others have decided to do.

The previous concept of active responsibility is captured by the deliberative STIT operator introduced in Sect. 2. This justifies the following abbreviation:

$$[i\ \mathsf{aresp}]\varphi \stackrel{\text{def}}{=} [i\ \mathsf{dstit}]\varphi$$

where $[i\ \mathsf{aresp}]\varphi$ has to be read "agent i is *actively* responsible for ensuring φ." The previous concept of passive responsibility is captured by the following abbreviation:

$$[i\ \mathsf{presp}]\varphi \stackrel{\text{def}}{=} \varphi \wedge \langle Agt \setminus \{i\}\ \mathsf{stit}\rangle [Agt\ \mathsf{stit}]\neg\varphi$$

where $[i\ \mathsf{presp}]\varphi$ has to be read "agent i is *passively* responsible for ensuring φ." According to this definition, agent i is passively responsible for ensuring φ if and only if, φ is actually true (φ) and i could have prevented φ from being true, regardless of what the others have decided to do ($\langle Agt \setminus \{i\}\ \mathsf{stit}\rangle [Agt\ \mathsf{stit}]\neg\varphi$).

We have already illustrated the meaning of the operator $[i\ \mathsf{dstit}]$. Let us illustrate the meaning of the passive responsibility operator $[i\ \mathsf{presp}]$ via the example of STIT model in Fig. 2. At world w_1 agent 2 is passively responsible for making p true. Indeed, p is true at w_1. Moreover, it is the case that, agent 2 could have prevented p from being true, regardless of what agent 1 has decided to do. The latter condition captures the fundamental counterfactual aspect of the concept of passive responsibility.

3.2 *Influence*

In Lorini and Sartor (2016), a STIT logic analysis of the concept of social influence is proposed. It starts from a general view about the way rational agents make choices. Specifically, the assumption is that an agent might have several choices or alternatives *available* defining her *choice set* at a given moment, and that what the agent does is determined by her *actual* choice, which is in turn determined by the agent's *choice context* including her preferences and beliefs and the composition of her choice set. The analysis of social influence expands this view by assuming that the agent's choice context determining the agent's actual choice might be determined by external causes. Specifically, the external conditions in which an agent finds herself or the other agents with whom the agent interacts may provide an input to the agent's decision-making process in such a way that a determinate action should follow. Specifically, influence consists in *determining* the voluntary action of an agent by modifying her *choice*

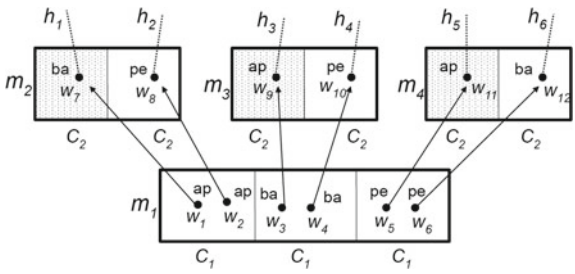


Fig. 4 Example of influence via choice restriction

context, so that a different choice becomes preferable to the influencee on comparison to what would be her preferred option without this modification. This may happen, for instance:

- by expanding the available choices (influence via choice set expansion), or
- by restricting the available choices (influence via choice set restriction) or
- by changing the payoffs associated with such choices, as when rewards or punishments are established (influence via payoff change).

The interesting aspect of STIT is that it is capable of: (i) capturing the temporal aspect of influence, namely the fact that the influencer's choice must precede the influencee's action,[5] and (ii) addressing the strategic aspect of influencing relationships through extensive form games.

To illustrate the concept of social influence, let us consider an example about influence via choice set restriction. The example is illustrated in Fig. 4. It represents a situation where there are three fruits on a table, an apple, a banana and a pear. The actions at issue consist in bringing about that the apple is eaten (*ap*), the banana is eaten (*ba*) or the pear is eaten (*pe*). Let us assume that agent 2 has certain preferences that remain constant along the tree structure. In particular, at all moments agent 2 prefers eating apples to bananas to pears. Let us also assume that 2 is rational, in the minimal sense that she acts in such a way as to achieve the outcome she prefers. Rational choices of agent 2 are depicted in grey. By choosing to eat the apple at w_1, 1 generates a situation where, given her preferences, 2 will necessarily eat the banana, rather than the pear. Indeed, although at moment m_2, 2 has two choices available, namely the choice of eating the banana and the choice of eating the pear, only the former is rational, in the sense of being compatible with 2's preferences. In this sense, by deciding to eat the apple at w_1 and removing this option from 2's choice set, 1 influences 2 to decide to eat the banana at w_7.

This example leads us to the following informal definition of social influence:

> An agent i influences another agent j to perform a certain (voluntary) action if and only if, i sees to it that that every rational choice of j will lead j to perform the action.

[5]The term "influencer" refers to the agent who exerts influence, whereas the term "influencee" refers to the agent being influenced.

In order to formalize the previous concept of social influence, in Lorini and Sartor (2016) STIT logic is extended by special "rational" STIT operators of the form $[i\ \mathsf{rdstit}]$. The formula $[i\ \mathsf{rdstit}]\varphi$ has to be read "if agent i's current action is the result of a rational choice of i, then i deliberately sees to it that φ." A minimal concept of rationality is adopted: it is assumed that the choices of an agent are ranked according to the agent's preferences, and an agent is rational as long as she implements her preferred choices. The $[i\ \mathsf{rdstit}]$ operator is interpreted relatively to STIT branching time structures, like the ones illustrated in Sect. 2. Specifically, the formula $[i\ \mathsf{rdstit}]\varphi$ is true at a certain world w if and only if, *if* the actual choice to which world w belongs is a rational choice of agent i *then*, at world w agent i deliberately sees to it that φ, in the sense of deliberative STIT discussed in Sect. 2. For example, at the world w_7 in Fig. 4, the formula $[2\ \mathsf{rdstit}]ba$ is true since the actual choice to which world w_7 belongs is a rational choice of agent 2 *and* at w_7 agent 2 deliberately sees to it that *ba* is the case.

To capture the idea of social influence, the following social influence operator based on the concept of deliberative STIT is introduced:

$$[i\ \mathsf{sinfl}\ j]\varphi \stackrel{\text{def}}{=} [i\ \mathsf{dstit}]\,\mathsf{X}\,[j\ \mathsf{rdstit}]\varphi$$

In other words, we shall say that an agent i influences another agent j to make φ true, denoted by $[i\ \mathsf{sinfl}\ j]\varphi$, if and only if i deliberately sees to it that if agent j's current choice is rational then j is going to deliberately see to it that φ. The reason why the operator $[i\ \mathsf{dstit}]$ is followed by the temporal operator X is that influence requires that the influencer's choice precedes the influencee's action. On the contrary, $[j\ \mathsf{rdstit}]$ is not required to be followed by X since in STIT the concept of action is simply captured by the deliberative STIT operator which does not necessarily need to be followed by temporal modalities. In order to illustrate the meaning of the influence operator, let us go back to the example of Fig. 4. Since agent 2 prefers eating bananas to pears, her only rational choice at moment m_2 is $\{w_7\}$. From this assumption, it follows that formula $[1\ \mathsf{sinfl}\ 2]ba$ is true at world w_1. Indeed, at world w_1 agent 1 deliberately sees to it that, in the next world, if agent 2's choice is rational then 2 deliberately sees to it that *ba* is the case.

Note that $[i\ \mathsf{sinfl}\ j]\varphi$ just says that the influencee i would realize φ is she were choosing rationally, but it does not assume that i chooses rationally, and therefore it does not entail that φ would be realized. The notion of *successful* influence also requires that in the next world along the actual history, the influencee chooses rationally, as specified by the following abbreviation:

$$[i\ \mathsf{succsinfl}\ j]\varphi \stackrel{\text{def}}{=} [i\ \mathsf{sinfl}\ j]\varphi \wedge \mathsf{X}\ \mathsf{ratCh}_i$$

where $[i\ \mathsf{succsinfl}\ j]\varphi$ has to be read "agent i *successfully* influences agent j to make φ true." The expression ratCh_i means that agent i's current choice is rational. It is an abbreviation, adopted for notational convenience, of $\neg[i\ \mathsf{rdstit}]\bot$, a formula that is satisfied only when i acts rationally in the current word.

This operator of successful influence clearly implies that in the next world along the actual history the influencee performs the action for which she has been influenced. This is expressed by the following valid formula:

$$[i \ \mathsf{succsinfl} \ j]\varphi \rightarrow \mathsf{X}\,[j \ \mathsf{dstit}]\varphi$$

In Lorini and Sartor (2016) is was also provided a complete axiomatization for STIT logic of social influence.

3.3 The Relationship Between Influence and Responsibility

The connection between the concept of influence and the concept of responsibility is tight and particularly relevant for legal theory. As legal theorists have emphasized (see, e.g., Hart and Honoré 1985; Kadish 1985), there exists a form of responsibility which consists in inducing another agent to violate a certain norm. In this case, the influencer becomes indirectly responsible for the violation of the norm, thereby being subject to a sanction. This captures the core of the idea of indirect (also called secondary or accomplice) responsibility in legal systems. In private law, there may be a "contributory liability" when someone with knowledge of the infringing activity, induces, causes, or materially contributes to the tort performed by another. In criminal law the idea of secondary responsibility concerns the connection between author of the crime, who performed the "actus reus" punished by the law, and his accomplices, who contributed to the performance of the "actus," without being part of it. For instance, the author of a robbery is the person, who breaks into a bank, threatens the clerks with a weapon and steals the money. Accomplices may have provided the weapons for the robbery, performed preliminary inspection of the places, or acted as lookout.

In Lorini and Sartor (2015), a formal theory based on STIT of the connection between influence and responsibility is provided. The relevance of such a theory for artificial intelligence (AI) lies in the possibility of exploiting it for automatic verification of secondary responsibility. Indeed, as highlighted above, the notion of influence is required to direct blame and sanctions not only against those who directly perform damaging acts, but also against those who have induced the authors to perform such acts. In the regulation of a society of artificial and/or human agents these responsibilities must also be introduced and checked, to effectively target cooperation aimed at socially obnoxious activities. This is clear for future society in which human agents will delegate tasks to autonomous agents and robots and will induce them to perform certain actions. Since responsibility can only be ascribed to humans, in case of violation of a norm by such artificial entities, humans agents will have to be sanctioned on the basis of their secondary responsibility for the violation.

As highlighted in Sect. 3.1, two basic forms of responsibility shall be distinguished, active responsibility and passive responsibility. Consequently, two forms of secondary responsibility shall be considered, namely active secondary responsibility

and passive secondary responsibility. Active secondary responsibility consists in an agent being actively responsible for the action of another agent and is captured by the following abbreviation:

$$[i \ \mathsf{saresp} \ j]\varphi \stackrel{\text{def}}{=} [i \ \mathsf{aresp}]\,\mathsf{X}\,[j \ \mathsf{rdstit}]\varphi \wedge \mathsf{X} \ \mathsf{ratCh}_i$$

where $[i \ \mathsf{saresp} \ j]\varphi$ has to be read "agent i is *actively secondarily* responsible for ensuring φ via agent j." This means that agent i is actively responsible for ensuring that every rational choice of agent j will result in φ true ($[i \ \mathsf{aresp}]\,\mathsf{X}\,[j \ \mathsf{rdstit}]\varphi$), under the condition that in the next world agent j will choose rationally ($\mathsf{X} \ \mathsf{ratCh}_i$). As the following valid formula highlights, it clearly coincides with the concept of successful influence defined in Sect. 3.2:

$$[i \ \mathsf{saresp} \ j]\varphi \leftrightarrow [i \ \mathsf{succsinfl} \ j]\varphi$$

Passive secondary responsibility consists in an agent being passively responsible for the action of another agent and is captured by the following abbreviation:

$$[i \ \mathsf{spresp} \ j]\varphi \stackrel{\text{def}}{=} [i \ \mathsf{presp}]\,\mathsf{X}\,[j \ \mathsf{rdstit}]\varphi \wedge \mathsf{X} \ \mathsf{ratCh}_i$$

where $[i \ \mathsf{spresp} \ j]\varphi$ has to be read "agent i is *passively secondarily* responsible for ensuring φ via agent j." This means that agent i is passively responsible for ensuring that every rational choice of agent j will result in φ true ($[i \ \mathsf{presp}]\,\mathsf{X}\,[j \ \mathsf{rdstit}]\varphi$), under the condition that in the next world agent j will choose rationally ($\mathsf{X} \ \mathsf{ratCh}_i$).

4 Deontic Extension

In Sect. 3.2, we have shown how STIT can be extended by a simple notion of rational (or preferred) choice. The idea is that at any given moment an agent i has a set of available choices. Only a subset of her available choices are rational and are denoted by the logical symbol ratCh_i. In particular, ratCh_i means that agent i's actual choice is rational.

In similar way, STIT can be extended by a simple notion of ideal choice denoted by the symbol idlCh_i. As for rational choices, the set of agent i's ideal choices at a given moment is a subset of agent i's available choices at this moment. An agent's ideal choices are those choices that are the best for the society, as they conform to its norms. Clearly, an agent's set of ideal choices does not necessarily coincide with her set of rational choices. For example, it might be ideal for an agent to help disadvantaged people in the society, even though it might be rational for her not to do it.

In the rest of the section, we explain how the selection of ideal choices out of the set of available choices is determined by the their ideality values. In particular,

ideal choices are those choices maximizing the ideality value. This utilitarian view of norms is commonly accepted in the deontic logic area to provide a formal semantics for normative concepts such as obligation "to be" (Føllesdal and Hilpinen 1971; Anderson 1957), obligation "to do" (Horty 2001; Kanger 1972; Sergot 1999) and dyadic or conditional obligation (Hansson 1969; Prakken and Sergot 1997). For example, the formal semantics for standard deontic logic (SDL) is based on the general concept of *ideality* according to which the set of ideal worlds (or situations) is a subset of the set of possible worlds (for a discussion of SDL, see Hilpinen 1982; Hilpinen and McNamara 2013). This is nothing but a special case of a more general formal semantics for deontic logic according to which possible worlds should be ranked in terms of their ideality values.

Ideality values of histories and ideality values of choices We enrich the STIT language with special atomic formulas of type $\mathsf{valCh}_{i,\geq x}$ and $\mathsf{valHis}_{\geq x}$, where i ranges over the set of agents Agt and x ranges over the set $\mathbb{N} \cup \{\omega\}$. $\mathbb{N}$ is the set of natural numbers and ω is the lowest transfinite ordinal number that is larger than all finite numbers in $\mathbb{N}$. The reason why we include ω is that we do not want to exclude histories which have assigned an *infinite* ideality value. The constant $\mathsf{valCh}_{i,\geq x}$ has to be read "agent i's actual choice has an ideality value at least x," while the constant $\mathsf{valHis}_{\geq x}$ has to be read "the actual history has an ideality value at least x."

The basic properties of these constants are captured by the following eight logical axioms:

(**ValCh**$_0$)	$\mathsf{valCh}_{i,\geq 0}$
(**ValHis**$_0$)	$\mathsf{valHis}_{\geq 0}$
(**ValCh**$_>$)	$\mathsf{valCh}_{i,\geq x} \rightarrow \mathsf{valCh}_{i,\geq y}$ if $x > y$
(**ValHis**$_>$)	$\mathsf{valHis}_{\geq x} \rightarrow \mathsf{valHis}_{\geq y}$ if $x > y$
(**ChDetValCh1**)	$\mathsf{valCh}_{i,\geq x} \rightarrow [i\ \mathsf{stit}]\mathsf{valCh}_{i,\geq x}$
(**ChDetValCh2**)	$\neg\mathsf{valCh}_{i,\geq x} \rightarrow [i\ \mathsf{stit}]\neg\mathsf{valCh}_{i,\geq x}$
(**TimeDetValHis1**)	$\mathsf{valHis}_{\geq x} \rightarrow (\mathsf{G}\ \mathsf{valHis}_{\geq x} \wedge \mathsf{H}\ \mathsf{valHis}_{\geq x})$
(**TimeDetValHis2**)	$\neg\mathsf{valHis}_{\geq x} \rightarrow (\mathsf{G}\neg\mathsf{valHis}_{\geq x} \wedge \mathsf{H}\neg\mathsf{valHis}_{\geq x})$

The first and second axioms state that every choice and every history have at least ideality value 0, as we assume that every choice and every history are identified with a non-negative integer.[6] The third and fourth axioms state that if $x > y$ and a choice/history has an ideality value at least x, then it has an ideality value at least y. According to the fifth and sixth axioms, the ideality value of a choice is choice-determinate. Finally, according to the seventh and eighth axioms, the ideality value of a history is history-determinate.

The following two abbreviations captures the exact ideality value of a choice and a history:

[6] For simplicity, we do not consider negative utilities.

$$\mathsf{valCh}_{i,x} \stackrel{\text{def}}{=} \mathsf{valCh}_{i,\geq x} \wedge \neg \mathsf{valCh}_{i,\geq x+1}$$
$$\mathsf{valHis}_{x} \stackrel{\text{def}}{=} \mathsf{valHis}_{\geq x} \wedge \neg \mathsf{valHis}_{\geq x+1}$$

where $\mathsf{valCh}_{i,x}$ and valHis_{x} have to be read, respectively, "agent i's actual choice has an ideality value equal to x," and valHis_{x} has to be read "the actual history has an ideality value equal to x." By convention, we assume that $\mathsf{valCh}_{i,\geq \omega+1}$ and $\mathsf{valHis}_{\geq \omega+1}$ are equivalent to $\bot$. Thus, $\mathsf{valCh}_{i,\omega} \stackrel{\text{def}}{=} \mathsf{valCh}_{i,\geq \omega}$ and $\mathsf{valHis}_{\omega} \stackrel{\text{def}}{=} \mathsf{valHis}_{\geq \omega}$.

Connection between ideality values of choices and ideal choices As we explained above, the atomic formula idlCh_i is aimed to identify agent i's ideal choices, that is, agent i's best choices according to the norms and standards of the society.

The following axiom captures the idea that at every moment an agent has at least one ideal choice:

$$(\textbf{AtLeastOneIdl}) \quad \Diamond \mathsf{idlCh}_i$$

It is justified by the assumption that there exists at least one choice that maximizes the ideality value.

The connection between ideal choices and ideality values of choices is captured by the following logical axiom:

$$(\textbf{ValChIdl}) \quad (\mathsf{idlCh}_i \wedge \mathsf{valCh}_{i,x}) \rightarrow \Box \neg \mathsf{valCh}_{i,\geq x+1}$$

The axiom means that if agent i's actual choice is an ideal choice and its ideality value is equal to x, then every available choice of agent i has an ideality value at most x.

Connection between ideality values of histories and ideality values of choices An important issue we have not yet addressed is the connection between ideality values of histories and ideality values of choices. There are different ways to represent this connection. One way is to assume that the ideality value of a choice corresponds to the *minimal* ideality value of a history passing through this choice. This is specified by the following logical axiom:

$$(\textbf{ValChValHis}_{min}) \quad \mathsf{valCh}_{i,x} \leftrightarrow (\langle i \ \mathsf{stit} \rangle \mathsf{valHis}_{x} \wedge [i \ \mathsf{stit}] \mathsf{valHis}_{\geq x})$$

Another way is to assume that the ideality value of a choice corresponds to the *maximal* ideality value of a history passing through this choice. This is specified by the following logical axiom:

$$(\textbf{ValChValHis}_{max}) \quad \mathsf{valCh}_{i,x} \leftrightarrow (\langle i \ \mathsf{stit} \rangle \mathsf{valHis}_{x} \wedge [i \ \mathsf{stit}] \neg \mathsf{valHis}_{\geq x+1})$$

The two Axioms ($\textbf{ValChValHis}_{min}$) and ($\textbf{ValChValHis}_{max}$) should be conceived as alternative criteria for selecting, among an agent's available choices, her ideal choices. In particular, Axiom ($\textbf{ValChValHis}_{min}$) together with Axiom (**ValChIdl**)

correspond to the *maxmin* criterion of selecting those choices whose worst possible outcome is better than the least possible outcome of all other available choices. Axiom (**ValChValHis**$_{max}$) together with Axiom (**ValChIdl**) correspond to the *maxmax* criterion of selecting those choices whose best possible outcome is better than the best possible outcome of all other available choices. Such criteria have been extensively studied in the area of qualitative decision theory (see, e.g., Brafman and Tennenholtz 1996, 2000; Goldszmidt and Pearl 1996).

Other criteria for selecting choices on the basis of ideality values of histories passing through them have been studied in the literature. For instance, Horty (2001) considers a selection criterion based on the concept of *dominance*. The idea is that an agent i's choice should be selected if and only if, there is no other choice of the agent that dominates it. It is said that agent i's choice A dominates choice B if and only if (i) for every possible choice of the other agents, choosing A is at least as good as choosing B, and (ii) there exists a possible choice of the others such choosing A is better than choosing B.

Obligation "to do" Following Horty (2001), we are finally able to formally define a concept of obligation "to do" (or "ought to do"):

$$\mathsf{Obg}_i\varphi \stackrel{\text{def}}{=} \Box[i\ \mathsf{idstit}]\varphi.$$

where:

$$[i\ \mathsf{idstit}]\varphi \stackrel{\text{def}}{=} \mathsf{idlCh}_i \rightarrow [i\ \mathsf{dstit}]\varphi$$

$\mathsf{Obg}_i\varphi$ has to be read "agent i is obliged to ensure φ," while $[i\ \mathsf{idstit}]\varphi$ has to be read "if agent i's current action is the result of an ideal choice of i, then i deliberately sees to it that φ." According to the previous definition, agent i has the obligation to ensure φ, denoted by $\mathsf{Obg}_i\varphi$, if and only if every ideal choice of agent i guarantees that agent i deliberately sees to it that φ.

Note that this concept of "ought to do" is essentially different from the concept of "ought to do" as formulated in the so-called Kanger and Lindhal's tradition (Lindahl 1977; Kanger 1972; Sergot 1999) by combining the "ought to be" operator of standard deontic logic (SDL) of the form O with the deliberative STIT operator $[i\ \mathsf{dstit}]$. As shown by Horty (2001), defining the normative sentence "agent i is obliged to ensure φ" by $\mathsf{O}[i\ \mathsf{dstit}]\varphi$ leads to counterintuitive consequences. For example, suppose agent i is a military general during a war. He has to decide whether (i) to bomb a enemy placement with the possibility destroying it without killing any civilian but with the potential risk of destroying it and killing civilians, or (ii) to refrain from bombing the enemy. The concept of "ought to do" represented by the combined operator $\mathsf{O}[i\ \mathsf{dstit}]$ tells us unambiguously that in this situation agent i has to bomb, as the only ideal situation is the one in which the enemy placement is destroyed and no civilian is killed. This conclusion is unsatisfactory, as it does not explain why the agent should unilaterally prefer one of the two options.

On the contrary, the previous concept of "ought to do" represented by the modal operator Obg_i is more flexible, as it tells us that the decision whether to bomb or not depends on the decision criterion adopted by the agent. In particular, it tells us that agent i should bomb if she adopts the *maxmax* criterion specified by Axiom (**ValChValHis**$_{max}$), whereas she should refrain from bombing if she adopts the *maxmin* criterion specified by Axiom (**ValChValHis**$_{min}$). Let us illustrate this in more detail. We present two different STIT models corresponding to these two different situations. Let proposition p denote the fact that "the enemy placement is destroyed" and proposition q denote the fact that "civilians are killed."

In the first model, represented in Fig. 5, ideality values of choices are determined via the maximality criterion formally represented by Axiom (**ValChValHis**$_{max}$) and ideal choices are determined via the corresponding *maxmax* criterion. In the second model, represented in Fig. 6, ideality values of choices are determined via the minimality criterion formally represented by Axiom (**ValChValHis**$_{min}$) and ideal choices are determined via the corresponding *maxmin* criterion.

Each history and each choice are identified with corresponding ideality values. In particular, history h_1 has an ideality value equal to 2 (formula valHis_2 is true at world w_1), history h_2 has an ideality value equal to 0 (formula valHis_0 is true at world w_2),

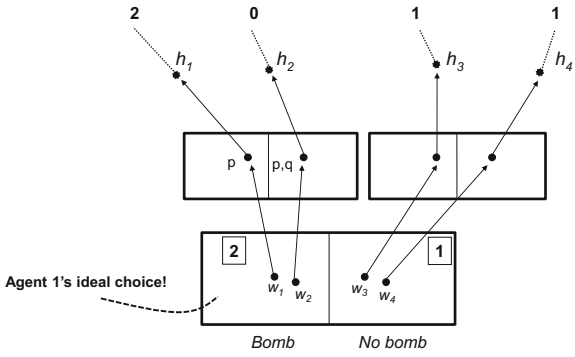


Fig. 5 Ideal choices determined via *maxmax* criterion

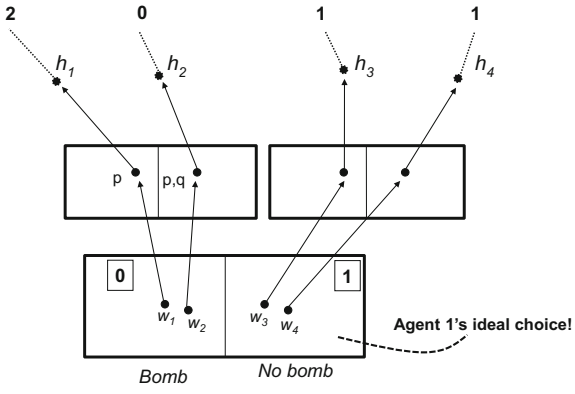


Fig. 6 Ideal choices determined via *maxmin* criterion

history h_3 has an ideality value equal to 1 (formula valHis_1 is true at world w_3), and history h_4 has an ideality value equal to 1 (formula valHis_1 is true at world w_4). We just assign arbitrary ideality values satisfying the following constraints: the situation in which the enemy placement is destroyed and no civilian is killed is better than the situation in which the enemy placement is not destroyed and no civilian is killed, and the situation in which the enemy placement is not destroyed and no civilian is killed is better than the situation in which the enemy placement is destroyed and civilians are killed.

Let us consider the model of Fig. 5. Agent 1's left choice in the initial moment containing world w_1 has an ideality value equal to 2 (formula $\mathsf{valCh}_{1,2}$ is true at worlds w_1 and w_2), while agent 1's right choice in the initial moment containing world w_1 has an ideality value equal to 1 (formula $\mathsf{valCh}_{1,1}$ is true at worlds w_3 and w_4). The ideality value of a choice corresponds to the *maximal* value of ideality of a history passing through it. It follows that idlCh_1 is true at worlds w_1 and w_2 and false at worlds w_3 and w_4. Indeed only agent 1's left choice is ideal. Consequently, agent 1 has the obligation to make p true (agent 1 has the obligation to destroy the enemy placement), as it is the case that p is true at every world included in agent 1's ideal choice. In particular, formulas $\mathsf{Obg}_1 p$ is at worlds w_1, w_2, w_3, and w_4. Furthermore, there exists a world in which p is false, which guarantees the satisfaction of the negative condition of the deliberative STIT formula $[1\ \mathsf{dstit}]p$.

On the contrary, in the model of Fig. 6, the ideality value of a choice corresponds to the *minimal* value of ideality of a history passing through it. It follows that idlCh_i is true at worlds w_3 and w_4 and false at worlds w_1 and w_2. Indeed, in this situation agent 1's right choice is the ideal one. Consequently, agent 1 has the obligation to make p false (agent 1 has the obligation to refrain from destroying the enemy placement), as it is the case that p is false at every world included in agent 1's ideal choice. Furthermore, there exists a world in which p is true, which guarantees the satisfaction of the negative condition of the deliberative STIT formula $[1\ \mathsf{dstit}]\neg p$.

5 Conclusion

Let us sum up what we have discussed so far. We have started with a concise presentation of the STIT formal language and semantics. Then, we have illustrated the use of STIT for the logical formalization of responsibility and influence, two concepts that are relevant for legal theory. Finally, we have presented a deontic extension of the STIT framework by a concept "ought to do," whose formal representation is based on the connection between the ideality value of a choice and the ideality value of a history passing through it.

An aspect we have not considered is the link between ideality values of histories and personal utilities of histories, where personal utilities are just the expressions of what the agents prefer. This is a fundamental component of norms of fairness and distributive justice. There are different ways of linking the two notions. For instance, Harsanyi's theory of fairness (Harsanyi 1982) provides support for an utilitarian

interpretation of ideality according to which the ideality value of a history coincides with the sum of the individual utilities of this history for the agents. An alternative to Harsanyi's utilitarian view of ideality is Rawls' view (1971). In response to Harsanyi, Rawls proposed the *maximin* criterion of making the least happy agent as happy as possible: the ideality value of a history coincides with the minimal utility of this history for the agents.

Another aspect we have not considered is the relationship between norms and agents' preferences. In Sect. 3.2, we have discussed an extension of STIT by a concept of rational (or preferred) choice. In real situations, an agent's rational choices may differ from her ideal choices. The aim of a normative system is to reduce the gap between the agents' rational choices and her ideal choices, by finding the appropriate balance of rewards and punishments.

We believe these two perspectives are promising, as they would complement the present study of the relationship between norm and action with an analysis of the relationship between norm and mind by considering (i) how what is ideal for the society may depend on the agents' personal utilities, and (ii) how norms may influence an agent's decision-making process.

References

Alchourrón, C.E., and E. Bulygin. 1971. *Normative systems*. Cambridge: Springer.

Alur, R., T. Henzinger, and O. Kupferman. 2002. Alternating-time temporal logic. *Journal of the ACM* 49: 672–713.

Anderson, A.R. 1957. The formal analysis of normative concepts. *American Sociological Review* 22: 9–17.

Aquinas, T. 1947. *Summa theologica*. Allen, Tex.: Benzinger Bros.

Balbiani, P., A. Herzig, and N. Troquard. 2013. Dynamic logic of propositional assignments: A well-behaved variant of PDL. In *Proceedings of the 2013 28th annual ACM/IEEE symposium on logic in computer science (LICS 2013)*, 143–152. Amsterdam: Morgan Kaufmann Publishers.

Belnap, N., M. Perloff, and M. Xu. 2001. *Facing the future: Agents and choices in our indeterminist world*. New York, N.Y.: Oxford University Press.

Belnap, N., and M. Perloff. 1988. Seeing to it that: A canonical form for agentives. *Theoria* 54: 175–199.

Bentham, J. [1872] 1970. *Of laws in general*. London: Athlone.

Blackburn, P., M. de Rijke, and Y. Venema. 2001. *Modal logic*. Cambridge: Cambridge University Press.

Bonzon, E., J. Lang, M.-C. Lagasquie-Schiex, and B. Zanuttini. 2006. Boolean games revisited. In *Proceedings of the 17th European conference on artificial intelligence*, ed. A.P.G. Brewka, S. Coradeschi, and P. Traverso, 265–269. ACM

Brafman, R.I., and M. Tennenholtz. 1996. On the foundations of qualitative decision theory. In *Proceedings of the thirteenth national conference on artificial intelligence (AAAI'96)*, 1291–1296, Palo Alto. California: AAAI Press.

Brafman, R.I., and M. Tennenholtz. 2000. An axiomatic treatment of three qualitative decision criteria. *Journal of the ACM* 47 (3): 452–482.

Broersen, J. 2011. Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic* 9 (2): 137–152.

Chellas, B.F. 1992. Time and modality in the logic of agency. *Studia Logica* 51 (3–4): 485–518.

Cicero, M.T. 1998. The laws. In *The republic and the laws*. Oxford: Oxford University Press.
Elgesem, D. 1997. The modal logic of agency. *Nordic Journal of Philosophical Logic* 2: 1–46.
Føllesdal, D., and R. Hilpinen. 1971. Deontic logic: An introduction. In *Deontic logic: Introductory and systematic reading*, ed. R. Hilpinen. Dordrecht: Reidel.
Gabbay, D.M., I. Hodkinson, and M.A. Reynolds. 1994. *Temporal logic: Mathematical foundations and computational*, vol. 1. Oxford: Clarendon Press.
Goldblatt, R. 1992. *Logics of time and computation*. Lecture Notes, Stanford, 2nd ed. California: CSLI Publications.
Goldszmidt, M., and J. Pearl. 1996. Qualitative probability for default reasoning, belief revision and causal modeling. *Artificial Intelligence* 84: 52–112.
Governatori, G., and A. Rotolo. 2005. On the axiomatisation of elgesem's logic of agency and ability. *Journal of Philosophical Logic* 34 (4): 403–431.
Grotius, H. [1625] 1925. *On the law of war and peace*, vol. 2. Oxford: Clarendon Press.
Hansson, B. 1969. An analysis of some deontic logics. *Mind* 3: 373–398.
Harel, D., D. K., and J. Tiuryn. 2000. *Dynamic logic*. Cambridge: MIT Press.
Harrenstein, P., J.-J. Meyer, W. van der Hoek, and C. Witteveen. 2001. Boolean games. In *Proceedings of the 8th conference on theoretical aspects of rationality and knowledge (TARK)*, 287–298. Amsterdam: Morgan Kaufmann Publishers.
Harsanyi, J. 1982. Morality and the theory of rational behaviour. In *Utilitarianism and beyond*, ed. A. Sen, and B. Williams. Cambridge: Cambridge University Press.
Hart, H.L.A., and T. Honoré. 1985. *Causation in the law*, 2nd ed. Oxford: Clarendon Press.
Hart, H.L.A. 1994. *The concept of law*, 2nd ed. Oxford: Oxford University Press.
Hilpinen, R., and P. McNamara. 2013. *Deontic logic: A historical survey and introduction*. London: College Publications.
Hilpinen, R. 1982. Deontic logic. In *The Blackwell guide to philosophical logic*, L. Goble ed. chap. 8, 159–182. Cambridge: Blackewell.
Horty, J.F., and N. Belnap. 1995. The deliberative STIT: A study of action, omission, and obligation. *Journal of Philosophical Logic* 24 (6): 583–644.
Horty, J.F. 2001. *Agency and deontic logic*. Oxford: Oxford University Press.
Jones, A.J., and M.J. Sergot. 1996. A formal characterisation of institutionalised power. *Logic Journal of the IGPL* 4: 429–445.
Kadish, S. 1985. Causation and complicity: A study in the interpretation of doctrine. *California Law Review* 73: 323–410.
Kanger, S. 1972. Law and logic. *Theoria* 38: 105–132.
Kelsen, H. 1967. *The pure theory of law*. Berkeley. California: University of California Press.
Leibniz, G.W. [1671] 1930. *Elementa Juris Naturalis*, vol. 1. Berlin: Akademie-Verlag.
Lindahl, L. 1977. *Position and change: A study in law and logic*. Dordrecht: Reidel.
Lorini, E., and G. Sartor. 2015. Influence and responsibility: A logical analysis. In *Proceedings of the twenty-eighth annual conference on legal knowledge and information systems (JURIX 2015)*. Frontiers in Artificial Intelligence and Applications, vol. 279, 51–60. Amsterdam: IOS Press.
Lorini, E., and G. Sartor. 2016. A STIT logic for reasoning about social influence. *Studia Logica* 104 (4): 773–812.
Lorini, E., D. Longin, and E. Mayor. 2014. A logical analysis of responsibility attribution : Emotions, individuals and collectives. *Journal of Logic and Computation* 24 (6): 1313–1339.
Lorini, E. 2013. Temporal STIT logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics* 23 (4): 372–399.
Meyer, J.-J.C. 1988. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic* 29 (1): 109–136.
Pauly, M. 2002. A modal logic for coalitional power in games. *Journal of Logic and Computation* 12 (1): 149–166.
Pörn, I. 1977. *Action theory and social science: Some formal models*. *synthese library*, vol. 120. Dordrech: Reidel.

Prakken, H., and M.J. Sergot. 1997. Dyadic deontic logics and contrary-to-duty obligations. In *Defeasible deontic logic*, ed. D. Nute, 223–262. Dordrecht: Kluwer.

Rawls, J. 1971. *A theory of justice*. Cambridge, Mass: Harvard University Press.

Raz, J. 1979. *The authority of law: Essays on law and morality*. Oxford: Clarendon Press.

Reynolds, M.A. 2003. An axiomatization of prior's ockhamist logic of historical necessity. In *Advances in modal logic*, vol. 4, 355–370. Amsterdam: King's College Publications.

Schmidt, R.A., D. Tishkovsky, and U. Hustadt. 2004. Interactions between knowledge, action and commitment within agent dynamic logic. *Studia Logica* 78 (3): 381–415.

Schwarzentruber, F. 2012. Complexity results of STIT fragments. *Studia Logica* 100 (5): 1001–1045.

Sergot, M. 1999. Normative positions. In *Norms, logics and information systems*, ed. P. McNamara, and H. Prakken, 289–308. Amsterdam: IOS Press.

Sergot, M. 2014. Some examples formulated in a 'seeing to it that' logic: Illustrations, observations, problems. In *Nuel Belnap on indeterminism and free action*, ed. T. Muller, 223–256. Berlin: Springer.

Tiomkin, M.L., and J.A. Makowsky. 1985. Propositional dynamic logic with local assignments. *Theoretical Computer Science* 36: 71–87.

van der Hoek, W., and M. Wooldridge. 2005. On the logic of cooperation and propositional control. *Artificial Intelligence* 164 (1–2): 81–119.

van der Meyden, R. 1996. The dynamic logic of permission. *Journal of Logic and Computation* 6 (3): 465–479.

von Kutschera, F. 1997. T $\times$ W completeness. *Journal of Philosophical Logic* 26 (3): 241–250.

Weinberger, O. 1998. *Alternative action theory. Simultaneously a critique of Georg Henrik von Wright's practical philosophy*. Berlin: Springer.

Wölf, S. 2002. Propositional Q-logic. *Journal of Philosophical Logic* 31: 387–414.

Wright, G.H.V. 1963. *Norm and action*. London: Routledge and Kegan.

Xu, M. 1998. Axioms for deliberative STIT. *Journal of Philosophical Logic* 27: 505–552.

Zanardo, A. 1996. Branching-time logic with quantification over branches: The point of view of modal logic. *Journal of Philosophical Logic* 67 (1): 143–166.

Of Norms

Jaap Hage

1 Terminology and Overview

Norms play a central role in practical reasoning, in law as well as in morality. An understanding of the nature of norms is therefore desirable for anyone who is theoretically engaged in practical reasoning, but such an understanding is not easy to achieve. The first challenge one encounters when trying to give an account of norms and their nature is that the terminology around norms is rich, to state it mildly. Von Wright starts his classic *Norm and Action* with the remark "The word 'norm' in English, and the corresponding word in other languages, is used in many senses and often with an unclear meaning" (Von Wright 1963, 1).

Ullmann-Margalit describes a social norm as "a prescribed guide for conduct or action which is generally complied with by the members of a society" (Ullmann-Margalit 1977, 12). However, she adds in a footnote that the term "norm" tends to be used by authors with a continental background, where authors with an Anglo-Saxon background prefer the terms "law" and "rule" (ibid.).

Kelsen writes at the very beginning of his posthumous study *Allgemeine Theorie der Normen* that the word "norm" denotes in the first place a command (*Gebot*), a prescription (*Vorschrift*), or an order (*Befehl*). He hastens to add, however, that ordering is not the only function of norms, but that empowering, allowing, and derogating are also functions of norms (Kelsen 1979, 1).

Although norms both play a role in law and in morality, the term "norm" has gained more popularity in the former field than in the latter. The term is also used more frequently in research from countries with a Roman law or Scandinavian background than in research from countries in the common law tradition. The reader may notice that these biases are reflected to some extent in the present contribution.

J. Hage (✉)
Faculty of Law, Maastricht University, Maastricht, The Netherlands
e-mail: jaap.hage@maastrichtuniversity.nl

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_5

Norms are both related to normativity and to rules. However, the notions of normativity and rules are as hard to pin down as the notion of a norm is. For example, the words "norm" and "rule" are sometimes used interchangeably for something that is normative. To clear out matters, this contribution will sharply distinguish between two oppositions: normative—non-normative and rule—description, and it will propose to use the term "norm" for normative rules only. In doing so, it must inevitably deviate to some extent from standard word use if there is something such as standard word use on these issues. The two distinctions will in the following be used as a framework for discussing not only norms as they are defined here, but also related phenomena that historically also have been called "norms."

The main distinctions will be explained in Sects. 2 and 3 of this contribution: Sect. 2 deals with the nature of normativity, while Sect. 3 focuses on the nature of rules. Norms are often opposed to facts, because the former would be normative while the latter would not. It will be argued that the distinction normative—non-normative—is not the proper basis to distinguish between norms and facts and to that purpose Sects. 4 and 5, respectively, discuss different kinds of facts and more in particular deontic facts such as the existence of duties and obligations. Section 6 returns to rules and distinguishes three different kinds of rules. The distinction is then used to identify norms in the strict sense defined here and to discuss the related phenomenon of rules that confer competence and other forms of legal status. The contribution will be summarized in Sect. 7.

2 Normativity

Norms[1] are basically used for two purposes. The one purpose is to evaluate states of affairs and acts, and the other is to guide human behavior.[2] In this contribution, the emphasis will be on the second function, but in order to avoid possible confusions it is useful to say a little here about the evaluative function of norms.

2.1 Ought-to-Be and Ought-to-Do

It has become customary to distinguish between norms of the ought-to-do type and the ought-to-be type.[3] A norm of the ought-to-do type tells us what to do, while a norm of the ought-to-be type informs us what should ideally be the case, without

[1] For now, the term "norm" will be used in a broad sense. The more specific use will be introduced in Sect. 6.

[2] It is also possible that norms guide behavior of non-human entities, such as computer programs and robots, but here we will not pay special attention to these possibilities.

[3] The distinction is already quite old. Von Wright (1963, 14) mentions in a footnote the work of Max Scheler (1954) and Nicolaï Hartmann (1962). In the *Handbook of Deontic Logic and Normative Systems*, Hilpinen and McNamara (2013, 97) refer to Castañeda, H.-N. Castañeda (1972).

specifying that somebody should do something. An example of an ought-to-do norm is the norm that house owners should clear away the snow from the pavement before their houses. This norm specifies that something should be done, and also indicates who should do it. An example of an ought-to-be norm would be that letters ought to be stamped. This norm does not specify that some action ought to be undertaken, let alone who is responsible for undertaking this action.[4]

While ought-to-be norms indicate what should ideally be the case, there is no logical connection between what ought to be done and what is ideally the case. If there is to be a connection between what ought to be done and what is ideal, this connection must be created by some perfectionist theory of practical reasoning, such as utilitarianism.

Since ought-to-be norms do not specify what ought to be done, they have no use in guiding behavior; they can only be used to evaluate states of affairs as right (in accordance with the norm) or wrong (in violation of the norm). Norms of the ought-to-do type, on the contrary, can both be used to guide behavior and to evaluate it. The norm that house owners should clear away the snow from the pavement before their houses directs house owners to clear away snow. Looking backward, it can be used to evaluate the snow-clearing act of a house owner as right. Looking forward, it can be used in justifying the judgment that it would be wrong if the house owners in a particular street would not clear away the snow. Notice that although both ought-to-be and ought-to-do norms can be used to evaluate, the former will be used to evaluate states of affairs, while the latter will be used to evaluate behavior.[5] Ought-to-be and ought-to-do norms have in common that they underlie binary evaluations in terms of right and wrong, and not grading evaluations in terms of better and worse.

Because the emphasis of this contribution will be on norms that guide behavior and since only norms of the ought-to-do type can guide behavior, ought-to-be norms will be left out of consideration from here on.

2.2 Influencing and Guiding Behavior

One function of norms is to guide human behavior. To guide behavior is not the same thing as to influence behavior, although there is an important connection between the two. If Adrian influences the behavior of Bernadette, Adrian does something that exerts a causal influence on what Bernadette does. For instance, because the traffic is heavy, Adrian clutches his six-year-old daughter Bernadette to prevent her from crossing the street. In this way, Adrian influences his daughter's behavior by making it impossible. Bernadette has no choice whether she will cross the street.

[4] It should be noted that, in particular in law, formulations that suggest an ought-to-be norm because they do not specify that something ought to be done can nevertheless stand for ought-to-do norms, because it is clear from the context who is responsible for bringing about the right state of affairs.

[5] Attempts to define ought-to-do norms in terms of states of affairs that ought to be the case (see Hilpinen and McNamara 2013, 97–112 for an overview) are in the eyes of the present author a major source of problems in formal deontic logic. See Hage (2001).

Another way to withhold Bernadette from crossing would be to warn her. If Adrian warns his daughter not to cross the street, he leaves the choice to Bernadette, but tries to influence the choice that she will make. This influence is a causal relation between performing the speech act of warning and the motivation of Bernadette.

It would be quite similar if Adrian got frightened when he saw his daughter approaching the busy street and yells "stop" to her. Again, this is a speech act aimed at exerting a causal influence on Bernadette's motivation. The type of the speech act might be described as "giving an order," but it should be noted that this order should not be seen as the imposition of a duty on Bernadette not to cross, but *merely* as an attempt to causally influence Bernadette's behavior.

Let us assume that Adrian, being Bernadette's father, has some authority over his daughter and that he can impose duties on her. Suppose moreover that Adrian exercises this power and forbids Bernadette to cross the street. The *ultimate* purpose of this prohibition is to withhold Bernadette from crossing the street, and in this respect the speech act of prohibiting is similar to the issuing of a warning or a mere order. However, there is an important difference between on the one hand the prohibition and on the other hand the order (and the warning, for that matter). The order is merely an attempt to causally influence Bernadette's behavior, with the causal relation being between the performance of the speech act and the motivation to refrain from crossing the street. The prohibition is a way to impose a duty upon Bernadette, and this duty also exists if Bernadette is not motivated to comply and crosses the street nevertheless. Moreover, the relation between the speech act of prohibiting and its consequence, the existence of a duty, is conventional, not causal. The causal influence between the prohibition and Bernadette's behavior, if it exists, goes via Bernadette's knowledge that she has a duty not to cross the street to her being motivated not to cross. Duties themselves do not motivate, but the awareness of an existing duty may.

The existence of a duty not to cross the street is a reason that applies to Bernadette—and in that sense is a reason for Bernadette—not to cross.[6] Reasons directly guide behavior by telling what is the right thing to do, and one can indirectly guide behavior by creating reasons. In our example, Adrian would guide the behavior of Bernadette indirectly by prohibiting her to cross the street. In doing this, Adrian creates a duty and therewith a reason for Bernadette to refrain from crossing and it is this reason that directly guides Bernadette's behavior. Notice that this guidance by the reason is not a causal influence. It is still possible that Bernadette ignores the reason and is not at all motivated by it. If that happens, the existence of the reason does not causally influence Bernadette. Still the guidance exists; it consists in an indication of what is the right or the good thing to do, and if Bernadette does not act on the reason most likely, she did something wrong.[7]

---

[6]Reason terminology has become dominant in ethical theory. See, for instance, Williams (1981), Alvarez (2010), and Broome (2013). In legal theory, it still lacks the popularity it deserves, despite the efforts of Raz (1975), Hage (1997, 2005), and Bertea (2009) to promote it.

[7]This is not the place to discuss the different functions of reasons for acting, and the distinction between guiding and explanatory reasons. However, it is worthwhile to point out that even if guiding

2.3 Guidance by Norms: The Second-Person Point of View

Norms are not the same things as reasons for action, but they are closely related. Suppose that the norm exists that pedestrians are not allowed to cross the street if the traffic lights for pedestrians are red. In that case, the fact that the traffic lights are red is a reason for Adrian not to cross the street. The norm generates reasons in all cases to which it applies, and in that sense the norm guides behavior in a general way.

The presence of reasons for action is typically expressed by the use of "normative" words, such as "shall," "should," "must," "obliged," "obligated," "duty," "obligation," "forbidden," "prohibited," "permitted," "allowed," and "ought." Some of these words express a situation in which not only an agent should do something, but also somebody (else) is entitled to claim from the agent that he[8] acts in a particular way. The entitlement to such a claim, which can sometimes be enforced, is characteristic for moral and legal norms.

Norms are characterized by what Darwall called "the second-person standpoint." Darwall described this second-person standpoint as "the perspective you and I take up when we make and acknowledge claims on one another's conduct and will" (Darwall 2006, p. 3). This standpoint is characteristic for both legal and moral norms, but seems to be lacking for many prudential reasons. For example, if Bertie is thirsty, she has a reason to take a drink, but—barring exceptional circumstances—nobody, not even she herself, can claim from her that she takes a drink. Law, morality, and prudence all provide agents with reasons for action, but law and morality are normative in a sense in which prudence typically is not.

The normativity of norms does not only involve that norms indicate *what* should be done, but also that they can indicate *how* things should be done. Assume by way of example that there is a norm prescribing that one should eat asparagus with one's fingers, rather than with fork and knife. Of course, there is no duty to eat asparagus, but somebody who eats asparagus should do so with his fingers. If he uses fork and knife, he does something that is wrong. These "how-to norms" should be distinguished from the "technical norms" that specify how something should be done in order to succeed in bringing about a particular result. Somebody who does not eat asparagus with his fingers still succeeds in eating asparagus, but somebody who tries to make a last will without witnesses will normally not succeed in making last will. We will return to "how-to norms" in Sect. 5.4.

---

reasons do not need to exert a causal influence on the person to whom they apply, the very notion of a guiding reason would not make sense if people in general would not be motivated by the awareness that a guiding reason applied to them. See also Sect. 5.1.

Legal philosophers will recognize the parallel with the relation between a legal system's efficacy and the validity of the rules that belong to the system. Validity cannot be derived from efficacy, but it makes little sense to speak of the validity of norms that belong to a system that is completely inefficacious (Kelsen 1945, 42).

[8]This contribution adheres to the convention that authors should use the pronouns for their own gender to refer to persons whose gender is not important for the argument or otherwise determined by the text.

2.4 Norms and Facts

Perhaps this is the right moment to briefly address the distinction that is traditionally made between norms and facts, a distinction that is often traced back to the work of Hume. A typical use of norms, in particular ought-to-do norms, is to evaluate acts. For example, the norm that house owners are to clean away the snow before their houses can be used to evaluate the cleaning as right or correct. Norms can only fulfill this function if they are somehow different from the acts that they are used to evaluate. The fact that house owners tend to clear their pavements does not coincide with the duty for house owners to do so, and it does not even have to be evidence for the existence of such a duty. If this distinction between the norm and the behavior that does or does not conform to this norm is meant by the distinction between norm and fact, it is obvious that the distinction between fact and norm exists.

However, it sometimes seems that the distinction between norm and fact is considered to be much more prominent, as if it were a major ontological divide between what is "out there" independent of human beings and their minds, and what is added by human minds to what "really" exists. This major ontological difference does not exist, or—if it is assumed nevertheless (see the discussion of "objective facts" in Sect. 4.1)—it is of limited importance. The distinction between deontic (normative) and non-deontic (non-normative)[9] is real, but has no major ontological relevance.[10] It is comparable to the epistemic difference between what is certainly the case and what is the case. However, in Sect. 3, a different distinction will be made which does have major ontological relevance. This is the threefold distinction between facts, constraints on facts, and descriptions of facts. This distinction, which has nothing to do with the distinction between deontic and non-deontic, is highly relevant for a proper understanding of rules and therefore also of norms as deontic rules.

3 Rules as Soft Constraints on Possible Worlds

Often the notion of a rule is connected to the guidance of behavior: Rules would indicate what we should do. On this interpretation, the meanings of the terms "rule" and "norm" would practically coincide. And yet there are "rules" whose primary function does not seem to be to guide behavior and which can therefore only be "followed" in a broad interpretation of that term. Examples would be rules that

[9]From here on, we will follow the custom among logicians and use the terms "deontic" and "non-deontic" for the distinction between normative and non-normative.

[10]The importance that is attached to the distinction between is and ought may be explained from the function of critical morality, that is, to evaluate existing moral practices critically. The social practice of critically moralizing can only exist if the fact that some norms are actually used does not count as sufficient evidence for the claim that these norms should be used. Ignoring the difference between the norms, people actually use and the norms people should use make critically moralizing impossible. So, there is a practical relevance to distinguishing between is and ought, but this relevance does not justify that the real difference is blown up to an ontological gap.

confer competences and rules which make that something also count as something else.[11] A proper account of the nature of rules would explain why some rules can be complied with and other rules cannot. Such an account would also clarify the nature of norms as a special kind of rules. The first step in providing such an account is to go into some detail concerning "directions of fit."

3.1 Directions of Fit

Perhaps the best way to introduce the distinction between directions of fit is by means of an example of Anscombe (1976, 56). Suppose that Elisabeth makes a shopping list, which she uses in the supermarket to put items in her trolley. A detective follows her and makes a list of everything that she puts in her trolley. After Elisabeth is finished, the list of the detective will be identical to her shopping list. However, the lists had different functions. If Elisabeth uses the list correctly, she places exactly those items in her trolley that are indicated on the list. Her behavior is to be adapted to what is on her list. In the case of the detective, it is just the other way round; the list should reflect Elisabeth's shopping behavior. The two different functions of the list with regard to Elisabeth's behavior reflect the two different directions of fit that we are looking for.

The two items involved in Anscombe's example are a linguistic one, the list of items and the world. The directions of fit distinction can also be applied to other items than purely linguistic ones, but let us focus on the purely linguistic case first.

The relation between language and the world goes in two directions. If the linguistic entities are to be adapted to the world, as when the detective writes down which groceries are in the trolley, the fashionable expression is "word-to-world direction of fit" (Searle 1979, 1–30) . If the world is to be adapted to the linguistic entities, as when Elisabeth puts those items in her trolley that are mentioned in her shopping list, the fashionable expression is "world-to-word direction of fit."

However, expressive as these expressions "world-to-word direction of fit" and "word-to-world direction of fit" may be, they are also difficult to keep apart. Therefore, it is proposed to use the different expressions, "down" and "up" (see Fig. 1). The basic idea is that descriptive sentences consist of words that aim to fit the world. The propositions expressed by them are true, and the speech acts in which they are used are successful in the sense of "truthful," if and only if the facts in the world correspond to ("fit"), what these propositions express. This is the up direction of fit.

For the down direction of fit, we must distinguish between three kinds. For all three kinds holds that somehow the facts in the world are adapted, in order to "fit" what is expressed by the words. One case is when the words function as a *directive*, as

[11] It is possible to construct these rules as elements of more complicated rules that do guide behavior, and that is why it was written that it is not their *primary* function to guide behavior. However, it is difficult to disagree with Hart (2012, 35–42), who wrote that the construction of such rules as parts of mandatory rules would be a distortion. Still there is a sense in which, for example, power-conferring rules can be followed, and in Sect. 4.2 an example will be discussed.

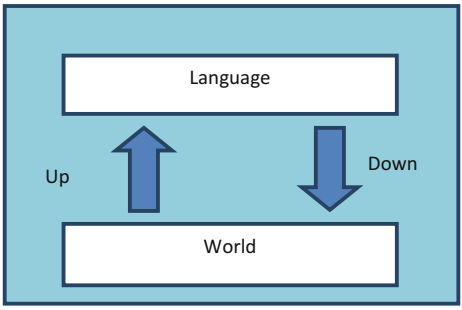


Fig. 1 Directions of fit

when Adrian shouts "Bernadette, stop!" when he fears that Bernadette will cross the busy street. This order aims at making its addressee stop, and if the order is successful in the sense of "efficacious," Bernadette will stop and the facts in the world fit the content of the order. In this case, the relation between the utterance of the order (the performance of the speech act) and the facts in the world is causal by nature. We might therefore speak of the "causal down direction of fit."

A second case concerns constitutive speech acts, such as "I hereby forbid you to cross the street." If such a prohibition is successful, the facts in the world come to match the content of the speech act and Bernadette has from that moment on the duty not to cross the street. In this case, the relation between the performance of the speech act and the facts in the world is constitutive by nature; the performance of the speech act constitutes the duty. We might therefore speak of the "constitutive down direction of fit."

Notice that this down direction of fit relates a speech act to a duty, not to the compliance with the duty. Efficacy is here the coming about of the duty which the speech act aimed to create. The duty itself can also be efficacious in the sense that it is complied with, but that would be an example of the causal down direction of fit.[12]

The third kind of down direction of fit concerns the effects of "constraints." Constraints will be discussed more extensively in Sect. 3.3, but here we will use one kind of constraint as example: the conceptual rule (rule of meaning) that makes that the bachelors are unmarried man. Given this rule, if somebody is a bachelor, he must be unmarried. This "must" depends on the conceptual rule that defines the relation between being a bachelor and being married. Given this rule, it cannot be otherwise than that a person who happens to be a bachelor is also unmarried.[13] The facts in

[12]Seemingly, the causal down direction of fit can also exist between duties and behavior, and not merely between speech acts and behavior. However, that would mean that non-material entities such as duties can exert causal influences, which do not sit well with our ideas about the nature of causation. It is therefore more coherent (with our views of causality) to say that the causal down direction of fit can exist between the belief that one has a duty, as realized by a brain state, and behavior.

[13]It may be disputed whether the fact that bachelor is unmarried depends on a conceptual rule and whether this conceptual rule does not depend itself on some ontological constraint (ontological nominalism or ontological realism). For the present purposes, this does not matter, however.

the world adapt themselves to the constraint, and that is what is meant by the down direction of fit of constraints.

In Sect. 6.2, we will see that the constitutive down direction of fit is a special case of the down direction of fit of constraints and more in particular of dynamic rules.

3.2 Possible Worlds

We are all familiar with the distinction between what the facts actually are and what the facts might have been. The sun is shining, but it might just as well have been raining. In Western Europe, there is peace, but there might have been a war. In the common law, judge-made law plays a paramount role, but its role might have been subordinate to statute-based law.

Logicians use possible worlds' terminology to deal with this distinction between what the facts are and what they might have been. They say, for instance, that in the actual world the sun is shining, but that in some other possible world it is raining. Intuitively, a possible world is a set of facts which makes some descriptive sentences true and others false. The set of facts that defines a possible world is complete in the sense that it determines for every non-modal descriptive sentence[14] whether it is true or false. The actual world is one of the many worlds that are possible, and in the actual world the sun is shining. However, in some other possible world, it is raining. That is another way of saying that although actually the sun is shining, it might have been raining.

Some things are necessarily the case. For example, five is necessarily bigger than three. If something is necessarily the case, there is no possible world in which it is not the case. In all possible worlds, five is bigger than three. And in all possible worlds, if Janet is either in Berlin or in London, and she is not in London, then she is in Berlin. Being necessary may just as well be circumscribed as being the case in all possible worlds. What is necessary is the case in all possible worlds, while what is impossible is not the case in any possible world. What is contingent is the case in some, but not in all possible worlds.

What makes that a world is possible, and how can we distinguish possible worlds from impossible ones? To answer these questions, we need to apply the idea of constraints to possible worlds. Constraints on possible worlds are limitations on which facts can go together and which facts exclude each other. Possible worlds satisfy these constraints, while impossible worlds violate one or more of them.[15]

[14]For ease of exposition, we will ignore here exceptional descriptive sentences, such as the sentences "The king of France is bald" and "This sentence is false." The clause "non-modal" was added to take into account that modal sentences which express necessity may be interpreted as dealing with more than one possible world.

[15]The metaphysics of possible worlds is the central topic of an anthology edited by Loux (1979). To the present authors' knowledge, however, the idea that possible worlds are relativized to sets of constraints is not treated in that anthology, nor in more recent overviews of the discussions about possible worlds, such as Menzel (2015).

Examples of such constraints are for instance physical laws. A physically possible world is a world that satisfies all physical laws, including the law that metals expand when heated. Since all physically possible worlds satisfy this constraint, in all these worlds pieces of metal expand when heated. The same thing, stated in terms of facts that go together, is that in all physically possible worlds the facts that *M* is a piece of metal and that *M* is heated go together with the fact that *M* expands. So, if we only look at physically possible worlds, it is necessarily the case that a piece of metal will expand if it is heated. It is also the case that a piece of metal would have expanded if, counterfactually, it would have been heated.

3.3 Constraints

Necessity and possibility are not absolute phenomena. Something is always necessary or possible relative to some set of constraints.[16] If all constraints, including those of logic, are left out of consideration, everything is possible, even what is logically impossible. Not all constraints are physical or logical. There are also conceptual constraints, such as the constraint that bachelors are unmarried males, that a rectangle has straight corners, and—perhaps more controversial—that gold is a metal. Some constraints seem to defy any category, such as the constraints—if they are actual constraints—that all colored objects have a surface, that an item cannot simultaneously be at two different places, that every event has a cause, and that causality does not operate backward in time.

We can try to imagine a world that is not constrained in any way. In that world, all facts are independent of each other, as are the truth values of propositions that purport to describe these facts. The truth of one proposition has no connection at all to the truth of any other proposition, and the relations between the truth values of all propositions would be like the relations between the truth values of atomic propositions in propositional logic. Although it may be hard to imagine, the proposition "It is now five o'clock and it is raining" might be true, while at the same time the proposition "It is raining" would be false. A world in which this is the case is logically impossible, but it is still a possible world if the logical constraints are left out of consideration.

A logically possible world is a possible world in which logical constraints determine (not necessarily exclusively), which combinations of facts always hold, and which other combinations of facts never occur. Exactly which combinations of facts are necessary or impossible depends on the precise nature of the logical constraints. One such a constraint is that a fact and the "opposite" of this fact cannot go together. For instance, if the fact that it is raining obtains in a logically possible world, then the fact that it is not raining does not obtain in that world. Or—to say the same thing in term of truth values of propositions—a logically possible world does not make

[16]The theory about constraints as exposed here has some remarkable similarities to the theory of modalities defended by Frändberg (typescript).

both the propositions "It is raining" and "It Is not raining" true. More in general, a logically possible world does not allow that a proposition and its negation are both true. Nor does it allow that a proposition and its negation are both false. These are both constraints on logically possible worlds.[17]

Constraints may seem mysterious entities, and the question is justified in which manner they exist. An attempt to explain necessity and possibility in terms of constraints seems a bit like the explanation of the sleep-inducing nature of opium by pointing to the *vis dormitiva*, the sleep-inducing power, of opium in Molière's play *The Imaginary Invalid*. That constraints somehow exist must be concluded from the fact that some things are necessary and others impossible. If we know that circles are necessarily round, we know something not only about actual circles, but also about the characteristics something would have if it were a circle, that is knowledge about possible circles. Knowledge about necessity is knowledge about hypothetical situations. This knowledge must be a priori (no dependent on sensory perception) and must be based on reasoning. The only feasible explanation that not only actual circles, but also possible circles are round is that the world is constrained in a manner that disallows non-round circles, and that this constraint also applies to the world that contains the possible circles about which we know that they must be round.

At first sight, it does not make much sense to search further for the nature of constraints and their mode of existence, but a little bit more can still be said. Take again the roundness of circles. This is often considered to be a conceptual truth. Unless one is a conceptual realist who assumes that concepts exist "out there," to be discovered by intelligent beings, one can assume that concepts are being created by human beings. In particular with regard to artificial concepts, such as "computer," it is plausible that they are human creations and could, at the time of their creation, be arbitrarily defined. It is still possible to modify the concept of a computer, to make it, for instance, include or exclude smartphones.[18] What a computer is, is a matter of convention, and the convention might have been slightly different from what it actually is. However, given the convention as it actually is, computers have some characteristics essentially and necessarily—for instance, having one or more processors—and other characteristics—for instance, their color—contingently and only possibly. The necessary characteristics of computers are based on a convention that functions as a constraint on what a computer can and cannot be.[19] Apparently at least some constraints are man-made, and rules belong to this category of man-made constraints.

---

[17] These constraints on logically possible worlds are typically represented in the semantics of logical theories by characteristics of the valuation function that assigns truth values to propositions. See, for instance, Navarro and Rodríguez (2014, 16).

[18] As a matter of fact, the concept of a "planet" has recently been redefined, taking away the status of a planet from the former planet Pluto. See https://en.wikipedia.org/wiki/IAU_definition_of_planet (last visited on December 24, 2015).

[19] This relation between conventions and the necessity based on them is explored a bit more in Hage (2013).

3.4 Rules as Soft Constraints

Why are rules a kind of constraints? Because they behave in many ways as other constraints. In the world in which a rule exists, the rule imposes itself on the facts of that world with the down direction of fit that other constraints also have. So, if some possible world contains the rule that thieves are punishable, then in this world thieves are punishable. Moreover, the rule also supports conditional and counterfactual judgments: If John had been a thief, he would have been punishable.

In a world that contains the rule that thieves are punishable, it is not merely a contingent matter of fact that thieves are punishable, but a necessary one, because being a thief makes one punishable. In this connection, something remarkable is the case. Rules allow for exceptions in the sense that sometimes the consequences of a rule do no hold, even though the conditions of the rule are satisfied. For instance, John, who is a thief, is also a minor and therefore the rule about the punishability of thieves cannot be applied to John. This possibility of exceptions seems hardly compatible with the necessary connection between being a thief and being punishable. And yet, the necessity and the exceptions have the same ground, which is that they are based on a constraint. Otherwise than descriptive sentences (see Sect. 4.6), constraints can have exceptions.[20] However, constraints also cover hypothetical and counterfactual situations, and that explains why judgments based on a constraint can express necessary relations such as the relation that thieves are necessarily punishable. Strangely, necessity and exceptions go hand in hand.

Rules have a lot in common with more traditional constraints such as the logical and physical ones, but there are also major differences. One such a difference is that rules only apply locally: The laws of one country are, for example, different from the laws of another country. The necessity of rule-based judgments seems therefore to be merely local necessity. This is different for logical and physical laws, which seem to have a universal scope of application.[21]

The scope of rules is not only limited in space, but also in time. Many rules can be created or derogated, and in that sense they differ from the more traditional constraints which somehow seem outside the scope of human manipulation. When the rule that thieves are punishable is introduced, suddenly all thieves become punishable. And when the rule is repealed again, the punishability of thieves disappears with the rule.

As a consequence of these differences, there can be some logically and physically possible worlds in which a particular rule exists, and other possible worlds in which the same rule does not exist. In a sense, it might be said that logical and physical constraints create necessities that are themselves necessary, while rules create

[20] Not only rules can have exceptions. There can also be exceptions to logical constraints (some descriptive sentences are not true or false) and to physical constraints (some physical laws are not applicable in extreme circumstances).

[21] This difference should not be overestimated, however. The geometrical law that the three corners of a triangle add up to 180° only holds for relatively small triangles and (which may be the same issue) for triangles in a flat plane. See also the discussion of the scope of physical laws in Toulmin (1953, 69 and 78).

contingent necessities. For this reason, rules will be categorized as "soft constraints," as opposed to the hard constraints that do not depend for their existence on human decision making or social practices.[22]

4 Kinds of Facts

If the notion of a fact is taken broadly as that aspect of reality which makes a true descriptive sentence true, there are many different kinds of facts: facts that exist "objectively," facts that depend on recognition, facts that are the results of rules or the use of reason, facts that are independent of all other facts, facts that "supervene" on other facts, "neutral" facts and facts involving evaluation, "inert" facts and facts that motivate or guide behavior. For a proper understanding of norms, the distinction between "inert" facts and facts that guide behavior may be the most important distinction between kinds of facts, but this distinction cannot be seen separate from many other kinds of distinctions between kinds of facts. Therefore, the present section and its subsections are devoted to a number of distinctions between kinds of facts. Their main purpose is to open up conceptual space for "deontic facts," facts that involve that something should, or ought to be the case, of that somebody should or ought (not) do something. These deontic facts are crucial to understand norms.

4.1 Objective Facts

Some facts seem to be objective. They include the facts that Mount Everest is a mountain, that it is higher than 8000 m, that there are lions and other kinds of animals, and that there are N suns, with N being some as yet unknown natural number. The objectivity of these facts lies in their being mind-independent, that is independent of anyone's beliefs, linguistic practices, conceptual schemes, and so on.

The idea that there are objective facts stems from a distinction that we make within our beliefs. To some beliefs, we ascribe a counterpart that somehow exists independently of what we humans think about them. This counterpart would consist of objective facts, and the facts mentioned above typically belong to this category. They are distinguished from other facts, which we take to depend on the human mind in some way. For example, many people take it that value judgments do not reflect an independently existing world of values, but rest on the way we humans evaluate things. No doubt evaluation is typically also based on objective characteristics of things, for instance the sharpness of the picture offered by the computer monitor, but these objective characteristics must be combined with a man-made standard to lead to the evaluation that this monitor is a good one. This dependence on a man-made

[22] This theme is elaborated in Hage (2015).

standard makes that the value judgment is not objective and that the fact expressed by it is not objective either. Objective facts are different, however. They are taken to obtain independently, and our knowledge, if it is objective, reflects these objective facts as they really are.

It may be argued that there are no objective facts in the sense of "objective" that is presently at stake. The reason is that every fact is the fact that …, where the dots are to be completed by some descriptive sentence. Facts depend on language, and since language is not mind-independent, facts are not mind-independent either, not even the "objective" ones. Many people would object against this conclusion because there must be something "out there" that precedes human categorization. That would be the "real" facts, and we humans try to develop concepts that fit this pre-linguistic substrate as well as possible. Whether such a pre-linguistic substrate really exists is an open question, but it certainly does not consist of the conceptualized reality in which we humans live. The assumption of a pre-linguistic substrate is the result of theorizing, not a precondition of it.

4.2 Brute Social Facts

When we are satisfied with a very coarse categorization, social facts may be described as facts which exist because the members of some group collectively recognize or accept them as existing. There are two variants on collective recognition. In the case of what we will call "brute social facts" the facts themselves are recognized by the members of some social group, while in the case of "rule-based facts" the facts are the result of some rule.[23] The facts based on rules that exist because of social recognition are perhaps better known as "institutional facts."

Brute social facts are the result of collective recognition. Important aspects of collective recognition are that sufficiently many and/or sufficiently important members of a social group believe the fact to be present, believe that the sufficiently many and/or sufficiently important other members also believe the fact to be present, and believe that these mutual beliefs constitute the believed fact.

Suppose that about 20 persons together make a foot trip to the top of a mountain. They believe that they are *as a group* walking to the mountain top, they believe that the others also believe that they are walking as a group to the top, and they all believe—minimally in the sense of not denying it when asked—that their mutual beliefs about acting together make that they are walking as a group to the mountain top, rather than as a set of individuals. In this case, the people in the group make a foot trip to the mountain top as a group. This is a brute social fact.

The above example deals with collective recognition of acting together, but collective recognition does not always deal with collective action. Suppose that one member of the group, say Henriette, utters strong opinions about which path to take

[23]The facts based on rules are perhaps better known as "institutional facts" (MacCormick and Weinberger 1986, 10).

to the top of the mountain and that most of the group members tend to act on these opinions. After having several times chosen a particular path because Henriette proposed to take it, most group members recognize the leading role of Henriette. They believe that Henriette has become the group leader that most other group members hold the same belief and that Henriette is the leader of the group because she is recognized as such by most group members. In this example, the brute social fact concerns the possession by Henriette of the status of leader of the group.

In the two above examples, recognition took the form of believing. However, sometimes mere believing does not suffice. If the leadership of Henriette in the group of mountain climbers has been sufficiently established, the group members may collectively recognize an order from Henriette as a reason for acting. Suppose that Henriette ordered Susan to walk on the back of the group to see to it that nobody stays behind. Then, Susan is considered to be obligated to walk on the back on the basis of collective recognition. This involves she is liable to be criticized by group members, including herself, if she does not walk on the back. Susan is obligated to walk on the back as a result of collective recognition, but Henriette's competence to create such a duty for Susan is also based on collective recognition. The group members collectively recognize an order from Henriette as creating a duty for the person who was ordered. They do this by collectively recognizing the duties that ensue from the orders. In combination, this amounts to the recognition of a power to create duties by giving orders. In its turn, the power actually exists by being recognized. This is an example of a fact—the existence of a power—that exists as the result of collective recognition, where the recognition does not consist in a belief, but in a complex set of dispositions to act.

4.3 Social Rules

If in our example about the mountain climbers the group members normally recognize the duties imposed by the leader of the group, whoever that may be, it may be said that the group has the rule that the group leader can impose duties. This rule exists through being recognized and therefore as a matter of social fact. The difference between having this rule and the recognition of the power of Henriette is the abstraction from the actual person having the power. When a power is not anymore ascribed to a particular person, but to a role—in this case the role of group leader—the acceptance of an ordinary social fact has become the acceptance of a social rule.

It is tempting to follow Hart (2012, 57) in assuming that the existence of a social rule involves the existence of a critical reflective attitude with regard to behavior covered by the rule. This characterization of social rules is quite adequate for rules that prescribe behavior, but less so for other kinds of rules such as power-conferring rules. A broader, and therefore more adequate, characterization of a social rule is that *a social rule exists within a group if sufficiently many (sufficiently important) members of the group recognize the consequences of the rule when the rule is applicable*. For a mandatory rule, this means that sufficiently many group members assume the

presence of a duty or obligation if the rule attaches this duty or obligation to an actual fact situation. If the duty or obligation applies to a specific group member, this recognition typically involves that the group member is motivated to comply with the rule. For a power-conferring rule, this means that sufficiently many group members recognize the power of a person to whom the rule conferred the power. This recognition typically consists in the recognition of the effects of the exercise of the power. In our example, this was illustrated by the group members recognizing the duty that Henriette imposed on Susan.

4.4 Rule-Based Facts

If the group of mountain climbers has the rule that its leader has the power to create duties for group members, the power of Henriette to impose the duty on Susan to walk on the back is an example of the application of this rule. Henriette's power exists because of the rule and does normally not require separate recognition of Susan's duty by the other group members. Because the group has this rule about the powers of its leader, Susan has the duty to walk behind as soon as Henriette has imposed that duty on her.

The fact that Susan has this duty exemplifies a rule-based fact. Rule-based facts are those facts which exist because they were attached by a rule to some other fact, including the occurrence of some event.[24] Law provides telling examples of rule-based facts. Suppose that the parents of Joan own the Blackacre Ranch. When they die, Joan inherits the Blackacre Ranch and becomes owner of the ranch, at the moment that her parents die, even though it may still take some time before people, including Joan herself, receive the information that this is the case and before people are in a position to recognize that Joan has become the owner.

When Joan becomes the owner of the ranch, she also becomes competent to mortgage the ranch and to transfer the ownership of the ranch to somebody else. Most likely, the ownership of the ranch also brings for Joan the duty to pay real estate taxes. All these facts obtain solely because of the application of rules to the existing facts. Rules can attach consequences to facts and to events, and these consequences are new facts, rule-based facts.

One kind of rule-based fact is the existence of another rule. Again, this phenomenon is particularly important in law where most rules exist because they were explicitly created. That the rule-creating events actually lead to new rules is because other rules attach the consequence that a rule exists to these events. A properly created rule immediately exists, even if its consequences do not receive any recognition yet. However, if the recognition of the rule consequences never occurs, or—in other words—if the rule is completely inefficacious, the rule stops existing. If the term "valid" is used for the existence of rules, this means that rules that belong to a legal

[24] In Sect. 4.6, these rule-based facts will be called "immediate rule-based facts," and they will be distinguished from "mediated rule-based facts."

system are valid if the system as a whole is efficacious; efficacy of the individual rule is in first instance not required. However, if an individual rule is or becomes inefficacious for a longer stretch of time (*desuetudo*), this may take away the rule's validity (Kelsen 1960, 215–219).

Also outside the law, rule-based facts play an important role. It becomes easier to recognize this when one sees that standards at the hand of which value judgments are given are also a kind of rules.[25] Suppose that a group uses the standard that a soccer match is good if the play is aggressive but not foul. Then, if some match has aggressive but not foul play, this match is good. This situation is not very different from that of the group that recognizes that Susan has the duty to walk on behind because the group leader said so. The fact that Susan has this duty is just as "real" as the fact that the soccer match is good. Obviously, the existence of evaluative facts depends on a presupposed standard, but this holds for all rule-based facts.

Still another example of rule-based facts is the facts expressed by the theorems of some branch of mathematics, systems of formal logic included. The theorems express facts, they are true propositions, but they derive their truth from the axioms and the semantic rules (the "valuation function") of the formal system to which they belong.

4.5 Creation and Derogation

Because the existence of rules is often rule-based facts, there is a risk that rules are confused with the events by means of which they were created. Legislation is then, for instance, seen as a collection of rules, rather than as a means to create rules.

A similar misunderstanding underlies the view that legal norms are a kind of commands.[26] Such a characterization of norms would be wrong because a command is a speech act and therefore a kind of event, while a norm is a rule and therefore not an event. The temptation to see norms as a kind of commands may be explained from the shared normativity of commands and norms; it almost seems as if the norms command to act in a particular way. However, a proper understanding of the mode of existence of norms should focus on norms being a kind of rules, rather than on the specific kind of rules that norms are.

Basically, the same kind of mistake is made if the existence of norms is somehow connected to a legislator.[27] Many norms have been created intentionally by some

[25] An important difference between these standards and, for instance, legal rules is that legal rules also generate exclusionary reasons (Raz 1975; Schauer 1991; Hage 1997), while evaluative standards typically do not. This difference has no fundamental consequences for the role of evaluative standards as underlying rule-based facts, however.

[26] Famously, Austin defined laws in his first lecture in *The Province of Justice Determined* (Austin 1954, 24) as commands which oblige persons generally to acts or forbearances of a class.

[27] Von Wright makes this mistake for a particular category of norms, the laws of the state. He calls such norms "prescriptions" and defines prescriptions as having their source in the will of a norm-authority (Von Wright 1963, 7). A similar mistake seems to be made by Alchourrón and Bulygin

authority, but being a norm, not even being a legal norm, is not the same as being created by a state authority or any other kind of authority.

The speech acts by means of which some norms are created should not be identified with the norms created by them or with any other kind of norm. *Mutatis mutandis* this also holds for speech acts by means of which norms are derogated or repealed. By passing a bill, a legislator can derogate existing norms, but that does not make the bill into a norm. The idea that there are derogating norms therefore rests on a mistake.[28]

4.6 Factual and Descriptive Counterparts of Rules

The rule that thieves are punishable makes it impossible that thieves are not punishable.[29] Or, to state the same thing affirmatively, the rule necessitates that thieves are punishable. The (existence of) the rule that car drivers must drive on the right makes that car drivers have the duty to drive on the right. And the rule that cars count as vehicles for the Traffic Law makes that cars are (count as) vehicles.

If some rule—or, more in general, a constraint—exists, this means that some general descriptive sentence will be true. This sentence has more or less the same formulation as the rule, but it is not the rule formulation but a sentence that aims to provide information about the facts. Since these facts obtain because of the constraint, this sentence will be true.

Such sentences are open generalizations. An open generalization is a generalization over potentially infinitely many items. Examples would be that pedestrians wear shoes (a false open generalization) and that atoms have a nucleus (a true open generalization). Examples of closed generalizations are that all desks in this classroom are brown and that all instances of the Olympic Games lasted less than four months. Open generalizations can have counter-instances and still be true. For example, the open generalization that birds can fly is true, notwithstanding the existence of ostriches. A counter-instance to a closed generalization falsifies this generalization.

A rule of thumb to distinguish open from closed generalizations is that a closed generalization requires the use of the word "all," or some equivalent, while this word can be left away in case of open generalizations. For example, the sentence "Desks in this classroom are brown" expresses an open generalization which is almost certainly false, even if all desks in the classroom happen to be brown. ("Happen" is another word that indicates a closed generalization.) The open generalization requires for its truth a law-like connection (a constraint) between being a desk in the classroom and

(1981), when they recognize an "expressive conception of norms," according to which norms are essentially commands.

[28]This mistake was made by Kelsen when he allowed the possibility of derogating norms (Kelsen 1960, 57, 1979, 1).

[29]Remember that this necessity is compatible with exceptions to rules. See Sect. 3.4.

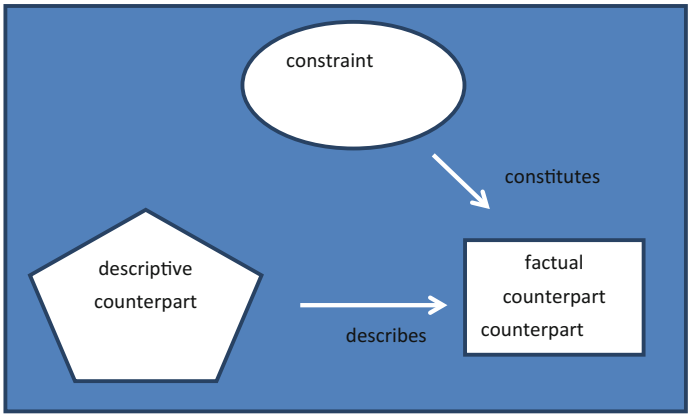


Fig. 2 Constraints, factual, and descriptive counterparts

being brown. Interestingly, it is precisely this law-like connection which makes that open generalizations can have counter-instances and be still true.[30]

The open generalizations that describe the effects of rules typically have the same formulation as the rule the effects of which they describe, and they are true because that rule exists. They describe facts that will be called the "factual counterpart" of the rule, and they may themselves be called the "descriptive counterparts" of rules. These descriptive counterparts of rules describe facts that are based on the existence of rules, but mediated by the rule-based facts that are immediately based on the rules.

Where rules impose themselves on the world by way of their down direction of fit, but are not true or false, the descriptive counterparts of rules are descriptive sentences, which are true or false, usually depending on the existence of the rules of which they are the counterpart. In schema (Fig. 2).

4.7 Norm-Propositions

The descriptive counterparts of norms have some resemblance with what are in the literature on deontic logic sometimes called "norm-propositions." Von Wright used the term "norm-proposition" for a proposition stating that a particular norm exists. Apart from the observation that Von Wright apparently saw norms as a kind

[30] See also Sect. 3.4. Because of the way open generalizations are often represented in formal logic, they have also become known under the misnomer "defeasible conditionals." Generalizations are not conditional sentences, even though they tend to be represented in predicate logic by means of conditionals. Moreover, open generalizations are true or false and not defeasible—but conclusions based on them may be defeasible (Hage 2005, 14). However, the truth conditions of open generalizations differ from those of closed generalizations, because the former are not necessarily falsified by counterexamples, while the latter are.

of entities that can exist, that is as a kind of logical individuals (see Sect. 6.1), these "norm-propositions" are not at all similar to descriptive counterparts of rules. Where norm-propositions in the sense of Von Wright talk about norms, descriptive counterparts talk about the subjects of norms (rules). For example, the descriptive counterparts of the norm "No vehicles in the park" are about vehicles, not about a norm.

Alchourrón and Bulygin (1981) and later Navarro and Rodríguez (2014) seem to follow Von Wright in calling statements about (the existence of) norms "norm-propositions," but they add that norm-propositions are statements about what is mandatory, prohibited or permitted relative to some set of norms. Since these statements are statements about prescribed, prohibited, or permitted states of affairs or actions, they are not statements about norms, so this identification seems to be based on a confusion. However, they are correct in pointing out that there are descriptive sentences, made true by existing norms (better: rules), stating that particular kinds of actions have a particular deontic status. Such statements describe if they deal with individual acts or with acts to be performed by a specific agent, rule-based deontic facts. Examples are the descriptive sentences "This killing was permitted" (based on the license to kill for secret agents) and "John must clear away the snow from the pavement before his house" (based on the rule prescribing house owners to clean away the snow before their houses). If the truth of these descriptive sentences is based on legal rules, these sentences would also express what Kelsen called "Rechtssätze" (Kelsen 1960, 57).

If such statements describe general prescriptions, prohibitions, or permissions, they are descriptive counterparts of rules. Examples would be "House-owners must clear away the snow from the pavement before his house" and "Secret agents are licensed to kill in the performance of her majesty's secret service." Kelsen called these general descriptive sentences "Rechtssätze" too (Kelsen 1960, 85), but also "legal rules" (Kelsen 1945, 45).

4.8 "Entailed" Norms

The recognition of descriptive counterparts of rules is important, because they are "ordinary" descriptive sentences to which deductive logic is applicable, whereas the applicability of deductive logic to rules, and to norms as a species of rules, is dubious since rules are from a logical point of view individuals (see Sect. 6.1). Many inferences which seem to have rules as their conclusions may well be interpreted as arguments with descriptive counterparts of rules as their conclusions. For example, the argument "Volkswagens count as cars. Cars owners must pay road tax. Therefore owners of Volkswagens must pay road tax" is dubious as an argument in which a rule

is derived from two other rules. As an argument in which two descriptive counterparts of rules are used to derive another open generalization, it is valid.[31]

The possibility to derive open generalizations from other open generalizations also explains the phenomenon of "deontic inheritance" (Hage 2001), "entailed norms" (Navarro and Rodríguez 2015), or "normative consequences" (Araszkiewicz and Pleszka 2015). An example would be that a prohibition for vehicles in the part would entail a prohibition for Volkswagens in the park. It is highly dubitable whether these entailed "norms" are rules that can be traced back to some official legal source. However, it is obvious that if vehicles are prohibited in the park, then—normally speaking—Volkswagens will be prohibited as special case of this general prohibition. The one deontic fact—vehicles being prohibited—encompasses the other—Volkswagens being prohibited—and there is no objection against deriving the proposition that expresses the latter fact from the proposition expressing the former fact. However, interpreting such a derivation as the derivation of one norm from, among others, another norm would be misguided (Hansen 2013).

5 Deontic Facts

5.1 Deontic Facts and Motivation

Norms are normative because they lead to duties or obligations.[32] Duties and obligations are entities that exist in time but not in space. They can be created and destroyed, and they can have all kinds of characteristics such as being a nuisance or being suitable to deal with societal problems.

The existence of a duty or an obligation is a deontic fact. It is a fact about an immaterial "thing," a duty, or an obligation, and it is deontic (normative) in the sense that it guides behavior. Suppose that Susan and Thera have concluded a labor contract. After that event, Susan has an obligation toward Thera to pay her a monthly salary, while Thera has an obligation toward Susan to work the afternoons of all weekdays. The sentences describing the existence of these obligations are true just like any other descriptive sentences. Facts that involve the existence of a duty or an obligation, and also some other kinds of facts, including the existence of permissions, may be called "deontic facts," after the convention that has arisen in logic to call logics that deal with duties, and obligations and with everything else that ought to be done "deontic logic."

---

[31] If the word "rule" is also used to denote open generalizations, there is no problem in deriving rules *in this sense* from other rules, *also in the sense of open generalizations*, and facts. It is this kind of reasoning about "rules" that seems to be at stake when MacCormick (1978, 100–108) writes about second-order justification of rules.

[32] The difference between duties and obligations as it is made here will be discussed in Sect. 5.2.

Sometimes the duties and obligations themselves, or their contents, are called "norms." As explained in Sect. 1, we adopted a different terminology here.

Even though the existence of a duty or an obligation is a fact, it is also deontic, normative, or behavior guiding, whatever you may want to call it. Although this is not the place to go into details with regard to the nature of normativity (however, see Sect. 2), it may nevertheless be useful to say a little about why duties (and obligations) are normative.[33] The normativity of duties lies in the connection, however remote that may sometimes be, between the existence of a duty and the motivation of persons to act in a particular way.[34] Typically, the acting person is the holder of the duty, and the behavior at issue is compliance with the duty. Agents tend to have a disposition to comply with their duties.[35] Although it is possible for an agent to recognize that he has a duty without being motivated at all, it is not possible that agents typically would not be motivated when they recognized to have duties. The reason is that if duty holders would not normally be motivated to comply with their duties, the very notion of a duty would not make sense.

The existence of a duty is not only based on the behavior or the disposition thereto of the duty holder; the behavior of agents in the environment of the duty holder is relevant too. This behavior consists of praise—in case the duty was complied with—or blame—in case the duty was violated, or—to use Hart's phrase—of a critical reflective attitude.

Sometimes the connection between a duty and the motivation to act is indirect. That is for instance the case with duties based on rules which exist themselves as a matter of rule-based fact, such as legal duties. Then, the disposition to comply with duties is the disposition of the addressees of the normative system as a whole to comply with duties based on the normative system. It is not the case anymore that every duty based on the system must lead to a disposition for compliance. Moreover, the efficacy of the system as a whole—because that is what we are talking about—must consist in recognition of the consequences of the system's rules, and since the rules are not always mandatory, the required efficacy is not always compliance with duties. It can, for instance, also be recognition of the power to make rules.

The connection between a deontic fact such as the existence of a duty and behavior may be quite complicated, but first, normativity cannot exist without such a connection, and second, there is nothing more to normativity than this connection. There is, for example, no such a thing as "binding force" apart from the disposition to motivate. That means that there is, for instance, no need to postulate the existence of a "norm"

[33]The following paragraphs only discuss duties, but *mutatis mutandis* the argument also applies to obligations.

[34]Sartor (2005, 454) seems to express the same idea when he characterizes obligations in terms of the intention to act on them.

[35]Human agents often critically evaluate the "duties" that they have according to a particular normative system, such as positive morality or positive law, and sometimes this evaluation leads to the conclusion that they should not comply with some "duty." However, such a refusal to comply with a "duty" which is not up to standard is often motivated by saying that the "duty" turned out not to be a "real" duty after all. In that case, the link between real duties and the motivation to act upon them remains intact. Obviously, much more can and needs to be said on this issue, but this is not the place to do so. Interested readers are referred to Hage (2013).

next to the rule-based facts that persons have certain duties and the constitutive rules on which these facts are based.[36]

5.2 *Duties and Obligations*

In normative discussions, words like "ought," "should," "must," "duty," and "obligation" are often used interchangeably. Although the meanings of these words in natural language are not fixed and overlapping, there exist different categories of deontic facts. The differences between these categories are important, although the words used to denote them are not. In the following, we will consider some distinctions and adopt particular words to denote the newly delineated categories. Let us start with some examples:

- Everybody has the duty not to steal, but normally there is no obligation to that effect.[37]
- *A* and *B* are under obligations toward each other, because they entered into a sales contract, but these obligations are not duties.
- From the fact that *P* is under an obligation or a duty to do something, it follows, *pro tanto*—that is, if only this reason is taken into account—that *P* ought to do it, but not the other way round.

The first distinction to be made is between duties and obligations. The existence of both a duty and of an obligation is a reason why somebody ought to do something, but neither the duty nor the obligation coincides with the fact that this person ought to do it. *A* duty is often connected to a role or status. It is, for instance, the duty of house owners to pay real estate tax and the duty of a mayor to maintain the public order in a municipality. All human beings[38] are under a duty not to kill other human beings. However, as our example of Bernadette and Adrian (Sect. 2.2) illustrated, it is possible to have a duty as the result of a command, and such a duty is not connected to a particular role or status.

Whereas duties are often connected to a particular status or role, an obligation is the outcome of an event and depends on that event having occurred. Typical examples of such obligation generating events are causing damage, making a promise, or contracting. Moreover, whereas a duty is not a duty with regard to

[36] Such a postulation seems to be made in the account that Navarro and Rodríguez give of the relation between norms and normative propositions (Navarro and Rodríguez 2014, 78).

The constitutive nature of the rules on which deontic facts are based is discussed in Sect. 6.5 of the present contribution.

[37] The term "obligation" derives the technical meaning that is proposed here from the civil law tradition, according to which an obligation is a particular kind of bond between a debtor and a creditor (for the historical roots of this word use, see Zimmermann 1996, 1). In the English literature, the difference between duties and obligations is not drawn sharply, possibly under the influence of the common law.

[38] Being a human being might be the most abstract status to which duties are assigned.

somebody in particular, obligations are always "directed," obligations toward somebody else.[39] This directedness of obligations still holds if this "somebody else" is (as yet) unknown, as when, for instance, a car was unlawfully damaged but the owner of the car is still unknown. Duties are not directed in this way.[40]

5.3 Being Obligated and Owing to Do Something

The term "obligated" will be used here as a term of art to denote the common denominator of duties and obligations: A person who is under a duty to do *B* is obligated to do *B*; a person who is under an obligation to do *B* is also obligated to do *B*. Being obligated is not directed. If *A* has contracted with *B* to pay him €100, then *A* has an obligation toward *B* to pay him €100, and *A* is also obligated to pay *B* €100, but *A* is not obligated *toward B* to pay *B* €100.

By now, we have encountered three normative concepts, "duty," "obligation," and "being obligated." They all differ from the normative concept that is often used as a catchall for all kinds of normativity, the concept of "ought." In connection with duties, obligations, and being obligated, the more relevant notion is ought-to-do. The word "ought" as defined here stands for the outcome of the interplay of one or more reasons for acting, a kind of aggregate of these reasons. Examples are the legal ought, as the aggregate of legal reasons for action, and the moral ought as the result of the aggregate of moral reasons.

An ought itself is not a reason for acting, but merely the outcome of one or more reasons. So, where the fact that *X* is under a duty to pay real estate tax is a reason why *X* ought to pay real estate tax, the fact that *X* ought to pay the tax is not a reason for paying it, although it *presupposes* the existence of such a reason (the duty, for example). An ought is comparable to being obligated in the sense that it abstracts from the precise reasons for acting, but nevertheless indicates (through presupposition) that such reasons exist. Where being obligated is tied to precisely one such a reason, owing to do something is based on a set of reasons, even though this set may contain one reason only. Being obligated can therefore be seen as a *pro tanto* ought.[41]

The difference between, for instance, an obligation and the ought based on it becomes clear if one considers what happens in case it is impossible to perform one's obligation. For instance, if Antony contracts with Giovanni to transfer his car to Giovanni, and if he also contracts with Guido to transfer his car to him, then

[39] For a logical discussion of these "directed obligations," see Herrestad and Krogh 1995.

[40] An example of a duty without a person toward whom the duty exists is the duty to stop for a traffic light, even if nobody is approaching. However, even if a duty mentions persons, e.g., the duty not to kill prisoners of war, this is not a duty toward these persons. Other persons can also address the duty holder about compliance with the duty. This is different for obligations, where typically only the right holders can demand compliance.

[41] The notion "prima facie ought" is more fashionable, but is strictly speaking an epistemic notion: If *A* prima facie ought to do *X*, then *for all we know A* ought to do *X*.

Antony both has an obligation toward Giovanni and toward Guido. It is impossible for Antony to comply with both obligations, and therefore[42] it is not the case that Antony both ought to transfer the car to Giovanni and to Guido. The law has a simple solution for such cases. Both obligations have an equal status (*paritas creditorum*). If Antony complies with his obligation to Giovanni, he must default on his obligation toward Guido, and—because the obligation still exists—Antony must compensate the damage of Guido. The question which obligation supersedes the other has a clear answer: neither one of the obligations *as such* supersedes. However, in determining what Antony ought to do, the reasons for acting have to be balanced. If the above account of the legal situation is correct, the outcome will be that Antony is legally permitted to deliver the car to any one of his creditors[43] and that he will have to financially compensate the other creditor.[44]

5.4 Permissions

Traditionally, permissions have been treated as the opposite of prohibitions. For example, if P is forbidden to do A, this means (is the same as) that P is not permitted to do A. However, it has turned out that the relation between on the one hand permissions and on the other hand the other deontic notions, such as "ought," "obligated," "duty," and "obligation," is not straightforward (Hansson 2013). Characteristic in this connection is the distinction, popularized by Von Wright (1963, 85–87), between weak and strong permissions. A weak permission would be nothing else than the absence of a prohibition, while a strong permission would involve a prohibition to interfere with an agent's freedom in a certain respect.

We will have a brief look at several possible interpretations of permission, and start with the possibility that act tokens, acts that have already been performed, were

[42] Although the principle "ought implies can" is in the eyes of the author not a logical constraint, there is from the moral and the legal point of view much to be said for it. In the law of obligations, for instance, impossibility is the main reason for assuming *force majeure*. That is why the principle is applied in the present argument.

[43] Whether Antony is also permitted not to deliver the car to anyone of his creditors depends on the legal system. In the common law, where "specific performance" is exception rather than the rule, Antony would be permitted to financially compensate both creditors rather than delivering the car to any one of them. In the civil law tradition, Antony would still be obligated to deliver the car to the creditor he does not compensate financially (Smits 2014, 194–202). This example illustrates in the first place that the relation between the existence of an obligation and what a debtor legally ought to do depends on the law, not on logic alone, and in the second place that it is useful to study the law from a comparative perspective to see the respective roles of law and logic. Where legal solutions differ, they cannot be a matter of logic.

[44] This account may not be correct for every legal system. In some systems, obligations to transfer a good do have a priority, with the older obligation superseding the more recent one. In those systems, the debtor legally ought to transfer the object to the oldest creditor. Also, this example illustrates that the relation between legal obligations and what an agent legally ought to do are in the first place governed by law.

permitted. In this brief discussion, the agents performing actions will mostly be left out of consideration and the talk will be about action types or act tokens that are, or are not, permitted. The evaluation of act tokens may be undertaken from several points of view, such as the legal and the moral point of view. Here, we confine ourselves to the legal point of view.

What does it mean that a particular act *token* was legally permitted? Basically, it means that this act, as performed by this agent, does not belong to a legally prohibited action type. Either there was no legal norm prohibiting this type of action, or in the concrete case there was an exception to the norm. Suppose that Ellen takes a break by making a walk on the lawn. Taking a bread was allowed, but walking on the lawn was not. Therefore, what Ellen did was not permitted, because her act can be subsumed under at least one prohibited action type. However, if Ellen would have received a special permission to walk on the lawn, her act was permitted, because the granted permission makes an exception to the general prohibition to walk on the lawn.

An action *type* is permitted if there is no norm which directly or indirectly prohibits that type of action. We will return to this distinction between direct and indirect prohibition soon, but first it is necessary to say something about default deontic status. Most of us live in a society where everything that has not been prohibited is permitted. Being permitted is the default deontic status of all action types, and it requires a prohibitive norm to change that status for a particular kind of action. Things might have been different, however. Logically, it would have been possible that everything that has not been permitted is prohibited. The reader should keep this in mind and be prepared to turn the following account of prohibited and forbidden action types around for the theoretical case that a society would have prohibition as its default deontic status.

An action type is directly prohibited if and only if there is a norm that explicitly prohibits that type of action. So, if the norm exists that prohibits lying, the action type lying is directly prohibited. If this norm does not exist, it might still be possible that there is a norm that explicitly forbids cheating. Then, cheating is directly prohibited. Suppose now, for the sake of argument, that lying involves cheating.[45] Then, barring exceptions, every instance of lying is also an instance of cheating. In that case, the direct prohibition of cheating makes that lying, the action type, is indirectly forbidden by the norm that prohibits cheating. Using this distinction between directly and indirectly forbidden action types, we can say that an action type is permitted if this type is neither directly nor indirectly prohibited.

Thus far we discussed permissions in the sense of absence of prohibition. An act token was permitted if it could not be classified as belonging to a prohibited type. An action type is permitted if there is no norm that directly forbids this type of action and if the performance of an act of this type does not involve the performance of a prohibited action. However, as was already pointed out by Von Wight, some

[45]The precise nature of this involves-relation which may hold between action types is crucially important in this connection. *A* first approximation would be that action type *A1* involves action type *A2* if necessarily every token of *A1* is also a token of *A2* (Hage 2001). It should be noted in this connection that the approximation presupposes that one act token can belong to more than one action type, and that the constraints that determine what counts as necessary remain unspecified.

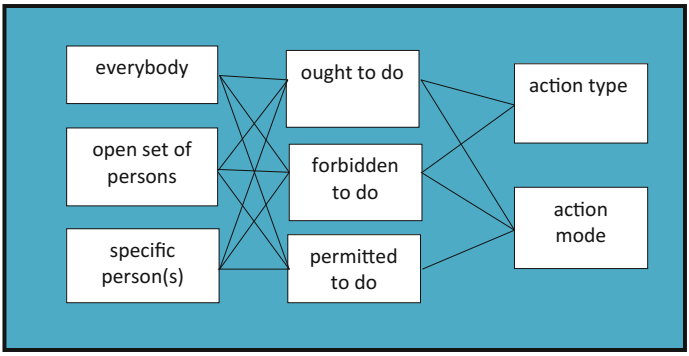


Fig. 3 Anatomy of ought-to-do

action types are explicitly permitted by a permissive norm. Such a permissive norm is typically—but not logically necessarily—connected to a freedom right. For example, the right to vote includes a permission to vote, and the freedom of expression includes a permission to utter one's opinions.[46] The permission that is included in some rights should not be confused with other elements that are also included in these rights. The idea, for instance, that strong permissions include Hohfeldian immunities—the legislator would, for instance, not have the power to forbid a citizen to vote—seems to be based on this mistake. In a right, a permission may be combined with an immunity, but this combination does not mean that the permission somehow includes the immunity. It is the right that includes both the permission and the immunity.

It is possible that an act token can be classified as belonging to two types, one type being explicitly permitted and the other type being forbidden. This would, for instance, be the case if it is forbidden to set people against each other, while it is permitted to express one's political opinions. Then, there is a norm conflict, which should be treated in the same way as other norm conflicts.

5.5 The Anatomy of Ought-to-Do

To understand the nature of deontic facts and of norms, it is useful to distinguish the elements of deontic facts. We will enumerate these elements for states of affairs of the ought-to-do type, but most of the discussion can mutatis mutandis be applied to duties and obligations as well (Fig. 3).[47]

[46]The inclusion of permissions—and also of competences—in rights is somewhat analogous to the involvement of one action type by another action type. An adequate theory about the nature of rights should include an elaboration of this includes-relation, but this is not the place to address this topic.

[47]The main difference is that whereas an ought and a duty contain three (or four) elements, an obligation is directed toward a creditor and therefore contains four (or five) elements, the extra element denoting the creditor.

An ought-to-do state of affairs involves that somebody is either permitted, required, or prohibited to do something, or to do something in a particular way, or at a particular time or place. An ought-to-do state of affairs consists of three or four elements, the deontic modality, the addressee, the act specification, and—occasionally—the specification of the act modality.

Take the following examples:

a. It is forbidden to murder.
b. Car drivers ought to carry a driver's license.
c. Leon is allowed to eat asparagus with his fingers.

In example a, everybody is an addressee; the modality is a prohibition, and the object of the deontic state of affairs is the performance of an action type (to murder).

In example b, the addressees are the members of the open class[48] of car drivers, the modality is an ought, and the object of this ought is the performance of an action type (carrying a driver's license).

In example c, the addressee is a single agent, the modality is a permission, and the object is an action mode (using one's fingers to eat asparagus).

It may be tempting to treat example b as expressing a conditional sentence: If somebody is a car driver, then he ought to carry a driver's license. This temptation is even strengthened if one considers how the deontic fact described in b can be used to argue why a particular car driver, say Lenny, ought to carry a driver's license. The following modus ponens style argument seems to do the job well: All car drivers ought to carry a driver's license; Lenny is a car driver; therefore, Lenny ought to carry a driver's license. Still, it seems a better idea to follow Von Wright (1963, 82) by distinguishing conditions under which a deontic fact obtains and the agents for which this deontic fact holds.

Notice, by the way, that this distinction presupposes that there are no conditional deontic facts, but only conditions for the existence of a deontic fact. If car drivers should place a warning sign before their cars if the car has broken down while it is dark, the deontic fact is that car drivers should put a warning sign before their cars, and this fact is present if the cars break down while it is dark. Of course, the sentence in which this relation between darkness and the duty to place a warning sign is expressed is itself conditional. However, that does not make the deontic fact conditional.

Only deontic facts where the addressee is a single agent (or a closed group of agents) concern the existence of a duty or obligation, and only these can be immediate rule-based facts. If it is a fact that car drivers ought to carry a driver's license, this is a fact only because every individual car driver, actual or merely hypothetical, ought to carry a driver's license. These individual oughts are most likely based on the rule (norm) that car drivers ought to carry a driver's license. A similar argument

[48]That the class is "open" means that the denoting expression refers to everybody who may happen to be a car driver and not merely to the fixed set of actual car drivers. The "openness" of the class makes that the deontic fact also deals with hypothetical car drivers, as in "If Thera would have been a car driver, she would have to carry a driver's license." See also the discussion of open generalizations in Sect. 4.6.

shows how it can be a fact that it is (for everybody) forbidden to murder. From this perspective, it can be seen that the sentences a and b are ambiguous: They may be read as rule formulations, but also as descriptions of deontic facts. In the latter case, they are the descriptive counterparts of the rules (see Sect. 4.6).

6 Of Norms and Other Rules

Given the facts that there is no fixed terminology concerning norms and that norms are closely related to rules and to normativity, it seems worthwhile to explore the idea that norms are rules—a kind of constraints—that lead to deontic facts. They lead in the first place to the existence of duties and obligations and—derived from these duties and obligations—to facts of the obligated—and ought type. We will explore that idea in this section, and to that purpose we start with distinguishing three kinds of rules, dynamic rules, fact-to-fact rules, and counts-as rules. Then, we consider the relevance of these kinds of rules for the constitution of deontic facts. This section is closed with some remarks on competence-conferring rules and other rules that confer status.

6.1 Rules as Individuals in the Logical Sense

Rule formulations such as "Thieves are liable to be punished" and "The Mayor of a municipality is competent to issue emergency regulations for that municipality" have, as far as their formulation is concerned, much in common with general statements. Moreover, rules can be used in rule-applying arguments which look like arguments of the modus ponens type. Nevertheless, there are important differences between rules and statements. Otherwise than statements, rules exist in time: They can be created and repealed (derogated; abolished). They can become outdated and can stop being used (*desuetudo*). Moreover, in contrast to statements which have the up direction of fit, they have the down direction of fit (of constraints). It is possible to predicate something of rules, such as in the sentence "This rule has been studied by legal historians for dozens of years." Rules can also be part of a relation, as can be stated in the sentence "The rule that thieves are liable to be punished exists longer than the rule that gives Mayors the competence to create emergency regulations."

Because of these latter reasons, there is much to be said for treating rules as a kind of things, rather than as statements describing what is the case. In the terminology of logic, such "things" are called individuals. If rules are from a logical point of view individuals, it is easy to see why rules as such cannot be parts of deductive logical derivations.[49] Deriving something from a rule would be comparable to deriving

[49]This does not exclude that rule-applying arguments are studied in logic. There are several ways to do so. One is to drop the demand that all elements of an argument are propositions. Second is

something from a chair. Norms do not figure in deductive arguments, but the reason is not that they are deontic or like imperatives (*pace* Jörgensen 1937/8), or that they have the down direction of fit of constraints,[50] but that they are from the logical perspective individuals.

6.2 Dynamic Rules

All rules connect facts to each other. These facts may be simultaneous, or they may succeed each other in time. The latter is the case with dynamic rules: They create new facts, or modify or take away existing facts as the consequence of the occurrence of an event.

Examples of events to which rules attach consequences are that John promised Richard to give him €100 and that Eloise was appointed as chair of the French Parliament. John's promise has the consequence that from the moment of the promise on John has the moral obligation to pay Richard €100. The appointment of Eloise as chair has as its legal consequence that from the starting point of the chair's new term on, Eloise will be the chair of the French Parliament. Other examples in which a dynamic rule attaches legal consequences to an event are that a Bill was passed, with as consequence the existence of new rules, or that Lionel committed theft, with as legal consequence that Lionel is liable to be punished.

Like all rules, dynamic rules have an element of generality. They apply to events of a particular kind and attach to these events facts of a particular kind. Dynamic rules may be conditional, in which case their consequence is only attached to the event under certain conditions. An example is the rule that if it is dark, the occurrence of a car accident obligates the drivers to place a sign on the road before to the cars.

If a juridical act or some other constitutive speech act is performed and a dynamic rule attaches consequences to this act, this is both a case of the constitutive down direction of fit and of the down direction of fit of constraints (see Sect. 3.1). The former focuses on the speech act by means of which the consequences were constituted; the latter focuses on the rule that attaches consequences to the performance of the act. The constitutive force of speech acts rests on the effects brought about by dynamic rules, and therefore the constitutive down direction of fit of constitutive acts is a special case of the more general down direction of fit of constraints and more in particular dynamic rules.

---

to allow entities without truth values as propositions. And third one is to use statements about the existence of rules as premises in rule-applying arguments. All of these options have consequences for the systems of logic that can be used to study rule-applying arguments that reach farther than accommodation for the defeasibility of rule-applying arguments.

[50] That the down direction of fit of constraints does not preclude them from being parts of arguments becomes clear from the example that if the world is constrained in such a way that Volkswagens are vehicles and that vehicles are not allowed in the park, the world is also constrained in the sense that Volkswagens are not allowed in the park. However, from the fact that the former two constraints exist as rules, it cannot be derived that the latter constrains also exist as a rule. See also Sect. 4.8.

6.3 Fact-to-Fact Rules

Where dynamic rules govern the succession of facts in time, static rules govern the coexistence of facts. One kind of static rules is fact-to-fact rules, rules which make that one kind of fact tend to go together with some other kind of fact, where the latter fact depends (supervenes) on the former. The relation between the kinds of facts is timeless, in the negative sense that the one kind of fact is not the occurrence of an event after which the second kind of fact comes into existence.[51]

A logical example of a fact-to-fact rule is that a conjunction is true if both conjuncts are true.

A moral example is the rule that spouses should be faithful to one another.

Legal examples of fact-to-fact rules are the rules that

1. The owner of a good is allowed to use this good.
2. The mayor of a municipality has the competence to issue emergency regulations for that municipality.
3. House owners must keep the pavement before their houses clean.
4. The king of Belgium is the commander in chief of the Belgian army.[52]

Characteristically, all the legal example rules attach consequences to the possession of a certain legal status. Important legal examples of fact-to-fact rules are rules that impose legal duties (example 3), rules that confer competences on people with a particular status (example 2), and rules that attach a specific status to the presence of some other status (example 4).

6.4 Counts-as Rules

The second kind of static rules that will be discussed here is *counts-as rules*. They have the structure: Individuals of type 1 count as individuals of type 2. These "individuals" may be human beings, as in the rule that the parents of a minor count as the minor's legal representatives. Often, however, the "individuals" that count as another kind of individual are events. For instance, under particular circumstances, causing a car accident counts as committing a tort, or offering money to another person counts as attempting to bribe an official.

Usually, counts-as rules are conditional, meaning that individuals of type 1 only count as individuals of type 2 if certain conditions are satisfied. An example from Dutch law (art. 3:84 of the Civil Code) would be the rule that the delivery of a good counts as the transfer of that good if the person who made the delivery was competent to transfer and if there was a valid title for the transfer.

---

[51] Notice that, this timeless relation between the conditions and the consequences of a fact-to-fact rule is compatible with the existence in time of the rule. Only as long as the rule exists, the condition facts and the conclusion facts go together in a timeless fashion.

[52] This last rule may also be interpreted as a counts-as rule.

Counts-as rules cannot create deontic facts by themselves. However, they often make that something counts as something to which a norm attaches deontic facts. Causing a car accident may count as a tort to which a rule of tort law attaches the obligation to pay damages. Being a person against whom serious objections exist counts as being a criminal suspect, a fact that gives police officers permission to arrest you.

6.5 Norms

To focus our discussion, we have defined norms as rules that constitute deontic facts. Let us focus our discussion even more, by assuming that there are only two kinds of basic deontic facts that matter for norms, that is duties, which are not directed toward a corresponding right holder, and obligations, which are directed in that way.

This stipulation contains two clauses that need justification. The first clause is that we confine ourselves here to *basic* deontic facts. That excludes facts of the types that somebody is obligated to do something or that somebody ought to do something. These latter facts are not basic, because they supervene on the existence of duties and obligations (see Sect. 5.2).

The second clause is that we confine ourselves to deontic facts that matter for norms. This has to do with the second-person perspective of norms (see Sect. 2.3): Norms justify claims for compliance. Mere requirements of practical rationality, such as the requirement that everybody who is thirsty should drink something, do not justify such claims. These requirements of practical rationality are not duties, let alone obligations, and the constraints underlying them, such as the constraint that somebody who is thirsty should drink, are typically not called norms. The practical relevance of this second point is mainly semantic, however.

We assume therefore that norms are rules that constitute duties or obligations. Examples of such norms are the norms that:

A. Nobody should steal (everybody has the duty not to steal).
B. Car drivers must (have the duty to) drive on the right-hand side of the road (perform acts in a particular way).
C. Paul must (has an obligation toward Patty to) compensate the damage of Patty.

As was to be expected, these norms have formulations that are identical to their descriptive counterparts. From the formulations, it is not possible to detect whether we are dealing with a norm or a descriptive sentence. The two major differences between norms and their descriptive counterparts are that norms have the down direction of fit of constraints, while their descriptive counterparts have the up direction of fit, and that norms are from the logical point of view individuals, and not part of language, while their descriptive counterparts are sentences and therefore part of language.

The account that was given above of norms emphasized the constitutive nature of norms: They create, rather than are, duties and obligations. This may look somewhat

strange at first sight, since the notion of a norm is more often associated with guidance of behavior than with the constitution of facts. Still this finding should not surprise if the idea is accepted that there can be facts that guide behavior. Rules that constitute such behavior guiding facts—that is, norms—are both constitutive and behavior guiding (regulative).

6.6 Competence-, Power-, and Other Status-Conferring Rules

The emphasis that is often placed on norms may draw our attention away from the rules that do not primarily aim at guiding our behavior. Still these other rules are crucially important for the functioning of more complex normative systems. To illustrate this, we will briefly pay attention to some rules that are not norms.

Dynamic rules attach new facts as legal consequences to the occurrence of some event. By performing some act that triggers the operation of a dynamic rule, an agent can bring about these legal consequences. For example, by committing a crime, a person can make himself liable to be punished, and by moving to a different municipality a citizen can change the amount of municipality tax he has to pay. The existence of a dynamic rule has a side effect that persons can do things which they could not have done without the presence of these rules. In that sense, they confer powers upon agents who through their acts are able to trigger the operation of dynamic rules and in that way bring about legal consequences.

The two examples given above of powers resulting from the existence of dynamic rules are not the most characteristic ones for law. The powers that can be exercised by performing a juridical act are much more characteristic. In this connection, a juridical act may be defined as an act that is aimed at bringing about legal consequences, to which these legal consequences are typically attached by a dynamic rule for the reason that this was the aim of the act.[53] Typical examples of juridical acts are contracting, making a last will or an association, legislating, pronouncing a judicial verdict, and granting a license.

An agent who performs a juridical act and thereby creates legal consequences must be competent to create these legal consequences by means of this kind of juridical act. This competence is assigned by a legal rule, although not necessarily an explicitly created one. For example, in order to be able to transfer the ownership of a piece of land, the transferor should be competent to alienate the land. This competence is

[53]The formulation "aimed at" has been chosen instead of the more natural sounding "performed with the intention to" to include acts that are performed without a conscious intention, such as acts performed by implemented computer programs. A public officer who signs a license without even reading it also performs a juridical act, and it should not be precluded by definition that a computer program that buys and sells securities thereby performs juridical acts. For this reason, the aim of an act should not be identified with the intention with which the act was performed. Aims are ascribed to acts, and (ascribed) intentions are merely factors that play a role in ascribing aims. Thanks go to Hester van der Kaaij for pointing out to me how important this innocuous-seeming difference between intention and aim is.

a legal status that is typically attached (by a fact-to-fact rule) to the ownership of the land. Being competent is a necessary condition for the successful performance of a juridical act and therefore also for the existence of a legal power that must be exercised through the performance of a juridical act. However, as we have seen from the two examples of the previous paragraph, legal powers, that is powers that exist because of the existence of legal rules, do not always require juridical acts for the exercise.

It is sometimes possible that somebody who lacks the competence for a particular juridical act nevertheless can succeed in bringing about the legal consequences of this act. For example, a public officer may succeed in providing somebody with a valid license, even though the officer lacked the competence to do so. Another example is that a non-owner may succeed in making somebody else the owner of a good. Both examples illustrate that legal consequences that could not be brought about through a valid juridical act may nevertheless occur for reasons of legal certainty. These examples illustrate that an agent can have the power to bring about legal consequences while lacking the competence to do so by means of a valid juridical act.

Rules that confer an agent the competence to perform a particular kind of juridical acts can both be dynamic and fact-to-fact rules. An example of the former is the dynamic rule that governs contracts and that by and large involves that the legal consequences which the contract partners intended to bring about will actually hold. If one contract partner wanted to make the other party competent to perform juridical acts in his name, that is to act as a legal representative, the effect of the contract is that this latter party has become competent to perform juridical acts in the name of the former contract partner. An example in which a competence is provided by a fact-to-fact rule is that the owner of a good is competent to alienate this good. The rule attaches the competence to the status of ownership.

Having a particular competence is a status assigned by a legal rule. There are many other examples of status assigned by legal rules. All legal counts-as rules assign a status to entities or events, such as the status of being a vehicle in the sense of the Road Act, being the president of Germany, being the commander in chief of the Belgian army, being a suspect in the sense of penal law, being the owner of a good, being wedded.[54] Very often norms attach deontic consequences to the presence of such a legal status.

7 Summary

In this contribution, an attempt was made to clarify the notion of a norm by elaborating the idea that norms are rules that lead to deontic consequences. The elaboration focused both on the nature of rules and on the nature of deontic facts.

[54] In a sense, even having a duty or an obligation can be seen as having a particular status, but this stretches the idea of legal status to its limits.

Rules, it was argued, are a kind of constraints on possible worlds. They determine which kinds of facts necessarily go together or cannot go together. Three kinds of rules were distinguished: dynamic rules which attach consequences to the occurrence of events; fact-to-fact rules which attach one fact to the presence of some other fact; and counts-as rules, which make that some things (often events) also count as something else. It was pointed out that the existence of a rule makes that some facts obtain: the factual counterparts of the rules. In this sense, all rules are constitutive. The descriptive sentences that express these facts, the descriptive counterparts of rules, are open generalizations, and they have often the same formulations as the rules from which they derive their truth.

By distinguishing between objective, brute social, and rule-based facts, an attempt was made to overcome resistance against the idea that facts might be normative and that there might be deontic facts. That something is mind-dependent does not exclude that it is a fact. Deontic facts are mind-dependent, because they are facts that tend—often in an indirect way—to induce a motivation to comply in agents to which they apply. Deontic facts are most often the result of the application of fact-to-fact rules (duties) or dynamic rules (obligations). A distinction was made between two kinds of basic deontic facts—the existence of duties and of obligations—and two kinds of supervening deontic facts: being obligated and owing to do something.

References

Alchourrón, C.E., and E. Bulygin. 1981. The expressive conception of norms. In *New studies in deontic logic*, ed. R. Hilpinen. Dordrecht: Reidel.

Alvarez, M. 2010. *Kinds of reasons. An essay in the philosophy of action*. Oxford: Oxford University Press.

Anscombe, G.E.M. 1976. *Intention*, 2nd ed. Oxford: Basil Blackwell.

Araszkiewicz, M., and K. Pleszka. 2015. The concept of normative consequence and legislative discourse. In *Logic in the theory and practice of lawmaking*, ed. M. Araszkiewicz, and K. Pleszka, 253–297. Cham: Springer.

Austin, J. 1954. *The province of jurisprudence determined*, ed. H.L.A. Hart. London: Weidenfeld and Nicholson. (1st ed. 1832).

Bertea, S. 2009. *The normative claim of law*. Oxford: Hart.

Broome, J. 2013. *Rationality through reasoning*. Chicester: Wiley Blackwell.

Castañeda, H.-N. 1972. On the semantics of ought-to-do. In *Semantics of natural language*, 2nd ed, ed. D. Davidson, and G. Harman, 675–694. Dordrecht: D. Reidel Publishing Company.

Darwall, S. 2006. *The second-person standpoint. Morality, respect and accountability*. Cambridge: Harvard University Press.

Frändberg, Å. Typescript. *The legal order. Studies in the foundations of juridical thinking.* (Typescript of a book in preparation).

Hage, J.C. 1997. *Reasoning with Rules*. Dordrecht: Kluwer.

Hage, J.C. 2001. Contrary to duty obligations. A study in legal ontology. In *Legal knowledge and information systems. JURIX 2001: The fourteenth annual conference*, eds. B. Verheij, A.R. Lodder, R.P. Loui and A.J. Muntjewerff, 89–102. Amsterdam: IOS Press.

Hage, J.C. 2005. *Studies in legal logic*. Dordrecht: Springer.

Hage, J.C. 2013. The deontic furniture of the world. In *The many faces of normativity*, ed. J. Stelmach, B. Brożek, and M. Hohol, 73–114. Kraków: Copernicus Press.

Hage, J.C. 2015. Separating rules from normativity. In *Problems of normativity, rules and rule-following*, ed. M. Araszkiewicz, P. Banaś, T. Gizbert-Studnicki, and K. Pleszka, 13–30. Cham: Springer.
Hansen, J. 2013. Imperative logic and its problems. In *Handbook of deontic logic and normative systems*, ed. D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, 137–191. London: College Publications.
Hansson, S.O. 2013. The varieties of permission. In *Handbook of deontic logic and normative systems*, ed. D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, 195–240. London: College Publications.
Hart, H.L.A. 2012. *The concept of law*, 3rd ed. Oxford: Oxford University Press. (1st ed. 1961).
Hartmann, N. 1962. *Ethik*, 4th ed. Berlin: De Gruyter.
Herrestad, H., and C. Krogh. 1995. Obligations directed from bearers to counterparties. In *Proceedings of the 5th international conference on artificial intelligence and law (ICAIL'95)*, 210–218. New York: ACM.
Hilpinen, R., and P. McNamara. 2013. Deontic logic: A historical survey and introduction. In *Handbook of deontic logic and normative systems*, ed. D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, 3–136. London: College Publications.
Jörgensen, J. 1937/8. Imperatives and logic. *Erkenntnis* 7: 288–296.
Kelsen, H. 1945. *General theory of law and state*. Cambridge, Mass.: Harvard University Press.
Kelsen, H. 1960. *Reine Rechtslehre*, 2nd ed. Wien: Franz Deuticke.
Kelsen, H. 1979. *Allgemeine Theorie der Normen*, eds. K. Ringhofer and R. Walter. Wien: Manzsche Verlags- und Universitatsbuchhandlung.
Loux, M.J. (ed.). 1979. *The possible and the actual. readings in the metaphysics of modality*. Ithaca, N.Y.: Cornell University Press.
MacCormick, N. 1978. *Legal reasoning and legal theory*. Oxford: Oxford University Press.
MacCormick, N., and O. Weinberger. 1986. *An institutional theory of law*. Dordrecht: Reidel.
Menzel, C. 2015. Possible worlds. In *The Stanford encyclopedia of philosophy,* ed. Edward N. Zalta. http://plato.stanford.edu/archives/sum2015/entries/possible-worlds/.
Navarro, P.E., and J.L. Rodríguez. 2014. *Deontic logic and legal systems*. Cambridge: Cambridge University Press.
Navarro, P.E., and J.L. Rodríguez. 2015. Entailed norms and the systematization of law. In *Logic in the theory and practice of lawmaking*, ed. M. Araszkiewicz, and K. Pleszka, 97–114. Cham: Springer.
Raz, J. 1975. *Practical reason and norms*. London: Hutchinson.
Sartor, G. 2005. *Legal reasoning, a cognitive approach to the law*. Dordrecht: Springer.
Schauer, F. 1991. *Playing by the rules*. Oxford: Clarendon Press.
Scheler, M. 1954. *Der Formalismus in der Ethik und die materiale Wertethik: Neuer Versuch der Grundlegung eines ethischen Personalismus*, 4th ed. Bern: A. Frankcke.
Searle, J. 1979. *Expression and meaning. Studies in the theory of speech acts*. Cambridge: Cambridge University Press.
Smits, J.M. 2014. *Contract law. A comparative introduction*. Cheltenham: Edward Elgar.
Toulmin, S. 1953. *The philosophy of science. An introduction*. New York, N.Y.: Harper Row.
Ullmann-Margalit, E. 1977. *The Emergence of Norms*. Oxford: Clarendon Press.
von Wright, G.H. 1963. *Norm and action. A logical enquiry*. London: Routledge and Kegan Paul.
Williams, B. 1981. Internal and external reasons. In Id., *Moral Luck*, 101–113. Cambridge: Cambridge University Press.
Zimmerman, R. 1996. *The law of obligations. Roman Foundations of the Civilian Tradition*. Oxford: Oxford University Press.

Values

Carla Bagnoli

1 Euthyphro Dilemma and Other Questions About Value

There are some fundamental problems concerning the meta-ethical status of values, their normative scope, and implications. These issues are importantly related, but it is useful to consider them separately. First, there is the *ontological* question: Is there value in the world? Are values part of the fabric of the universe, or else artifacts, projections of the mind, products of social conventions? In the platonic dialogue, Euthyphro asks whether the gods love the good because it is good or else the good is good because the gods love it (Plato 1991 6e–9e, 9d–11d). This is a fundamental dilemma about the nature and the ontological status of values, which divides realists and anti-realists. If what is good is good because gods love it, then the good is relative to them and dependent on their attitude. Vice versa, if gods love the good because it is good, it must be a real feature of the world, independent of what the gods happen to like or dislike. In the former case, values are subjective and their aspiration to objectivity seems problematic. In the latter case, the ontological status of values seems firmer and grounds the aspiration of value judgments to objectivity.

However, if values enjoy such an independent ontological status, it becomes unclear how they relate to other factual elements of the fabric of the world, but also how they matter to evaluators. This is a question that applies also in the case of the Platonic gods, but it becomes particularly interesting in the case of standard evaluators. How do objective values affect the life of such evaluators, presumably limited in their epistemic and practical rationality? This is the core issue of meta-ethics that aims to vindicate both the aspiration to objectivity and the practical relevance of

C. Bagnoli (✉)
Dipartimento di Studi Linguistici e Culturali, Università di Modena e Reggio Emilia, Modena, Italy
e-mail: carla.bagnoli@unimore.it; carla.bagnoli@gmail.com

C. Bagnoli
Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Oslo, Norway

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_6

value judgments. The tension between these two aspirations to objectivity and to practicality is so central that it has been dubbed "*the* moral problem" (Smith 1994). However, as we shall see, this apparent tension can be solved by understanding objectivity in a way that does not privilege ontology. After all, an objective value is also expected to exercise authority and bear some relation to our practical lives. Thus, the key problem is how to understand value objectivity.

Some of the problems generated by the search for objectivity are related to a well-established dichotomy between *matters of fact* and *matters of value*. Is this dichotomy to be taken for granted? The axiological status of subjects, agents, and objects crucially depends on what happens in the natural world. Differently than other forms of dependence, the dependence of values on facts is a conceptual truth. That is, it appears to be true a priori that the axiological status on an object cannot vary unless something in its natural makeup varies (Moore 2010; Hare 1963). In other words, two states or affairs or objects that are identical in their natural makeup must have identical axiological status, hence be evaluated similarly. How to understand and account for this sort of co-variance between factual and evaluative aspects is a problem that any theory of value should address. Reductive naturalists insist that the co-variance is implicated by a relation of identity; moral facts vary according to natural facts just because they are reducible to, or identical with, them. By contrast, anti-naturalists hold that moral facts or moral properties supervene upon natural facts and properties, but they are not reducible to them. An intermediate view holds that moral and natural concepts are not identical, but they share the same natural ontology. This view has the advantage of preserving the independence of the evaluative conceptual apparatus without committing to a special ontology. By contrast, an extreme position requires that ethical concepts be eliminated because they are not supported by any ontology.

A second, related, question concerns the *epistemology of value*. Can there be knowledge of right and wrong, good and evil? If there is, how is it accessible to us? Cognitivism and non-cognitivists offer opposite answers. If we can acquire knowledge of values, is it by education or by exercising and developing natural dispositions? This question admits of large varieties of positions. We can distinguish two main approaches, the rationalist and the sentimentalist. On the first approach, value is accessed by exercising rationality. The two main varieties of this approach are the rationalist intuitionism and constructivism, Rational intuitionism claims that we know right from wrong because reason recognizes it (Moore 2004, 2010; Ross 1930 ; Audi 2001; Dancy 2000, 2004; Skorupski 2010 ; Parfit 2011). By contrast, constructivism holds that practical knowledge is provided by engaging in practical reasoning (O'Neill 1989; Korsgaard 1996; Scanlon 1998, 2014; Bagnoli 2013; Wong 2008). Instead, on the sentimentalist approach, sensibility is the source of value and of moral distinctions (Blackburn 1998; Nichols 2004; Prinz 2007; Schroeder 2010, 2011). On this view, reasoning is driven by emotions and we have knowledge of right and wrong because we exercise virtues, which are a natural endowment of our sensibility (Aristotle 1984;Baier 1985, 1994; Foot 2000; Geach 1956; Hurka 2000; MacIntyre 2007, 2008; Slote 1992, 2010).

A third, related issue concerns the *semantic and logical* status of value judgments and the logical grammar of evaluative concepts. Are value judgments assertions

about properties or states of affairs, or are they akin to prescriptions or expressions of normative states? Do evaluative concepts represent properties or facts of the world, or are they expressive of mental states? If evaluative judgments are not assertions about states of affairs and evaluative concepts do not represent properties and facts of the world, how are they anchored to reality? If they are not anchored to reality, how can they be responsive to facts and subjected to rational criteria of justification and revision? Realism is the view that ethical judgments are assertions about properties or facts of the world, and thus behave exactly like other sorts of factual assertions. Anti-realism denies all the above. Emotivism holds that evaluations are merely expressions of mental states, hence cannot be true or false (Ayer 1936; Stevenson 1937, 1963, 1979). The major problem of emotivism is that it seems to fail to account for central linguistic phenomena such as embedding moral and other evaluative expressions within conditionals (Geach 1960). The major problem of realism is that it bases its claims on truth of a special ontology, which is problematically related to natural ontology and may seem to require a special faculty of knowledge (Blackburn 1993; Mackie 1977, 44ff; Wong 1986).

These positions identify the two extremes of a large spectrum of more nuanced and hybrid positions. Universal prescriptivism is a similar view, but it claims that evaluative judgments have logical and semantic properties that generate practical reasoning (Hare 1952, 1963, 1981). Normative expressivism is the view that judgments of value are expressive of emotions of normative states, which are themselves governed by logical norms, in ways that seem to avoid Geach's problem about embedded contexts (Gibbard and Macintyre 1995). Quasi-realism holds that evaluative judgments only resemble assertions but they are not. They exhibit some sort of objectivity, but they have to earn their right to truth (Blackburn 1984). Error theory holds that value judgments are like assertions, but they are all false because they do not represent any real property or fact of the world. Both quasi-realism and error theory involve a systematic error on the side of the evaluator; hence, they are committed to a projectivist account of moral phenomenology, which is supposed to explain how it happens that people are trapped in this illusion, and how value judgments can guide them nonetheless. By contrast, cognitivist irrealism holds that evaluative judgments have cognitive import, but they do not represent a special sector of reality (Skorupski 1993, 2010). Kantian constructivism can be plausibly defended as a form of cognitive irrealism (Bagnoli 2013).

Finally, a large cluster of questions concerns the *normative authority* of values. Do values produce normative reasons? Do such reasons apply universally, across all evaluators, in all relevant circumstances? Does the adoption of values commit the rational agent to act for the pursuit of value? At least some values entail normative reasons. For instance, if Spartacus values equality, then he has a reason to support policies that promote equality. Such reasons may not be overriding; that is, they can be defeated by stronger reasons. Brutus values friendship, and he has a reason to love and protect his friend Caesar. Yet Brutus decides to participate in Caesar's assassination, not because he loved Caesar less, but because he loved Rome more. In this case, the value of civic friendship trumps the value of personal friendship.

Does civic friendship always trump personal friendship? Do patriotic values generate normative reasons that apply to all of us?

Arguably, moral values, such as the value of humanity, produce reasons that trump all other kinds of normative reasons. Especially for such values, the question arises about the scope of their applicability. According to some ethical theories, moral values are universally authoritative and give everyone normative reasons. Rationalist theories and Kantian theories hold that moral obligations are requirements of reason, hence applicable to all rational agents (Kant 1997; Audi 2005). They claim that the moral values are universally compelling exactly because they are required by reason (Parfit 2006).

However, a further question is whether the requirements of reason are universally compelling (Broome 2013; Kolodny 2005). Some are skeptical that reason may be universally binding not because human agents are defective but because rationality produces merely conditional requirements. Relativist theories hold that moral values are rooted in particular traditions and that their authority is local (MacIntyre 1988; Wong 2006). Consequently, values generate reasons that are binding and authoritative only for the members of specific communities, and even though they claim universal authority, such a claim is misplaced. The traditional objection against relativism is that of incoherence (Lyons 1976), but Harman (1977) attempts to avoid the problem by distinguishing between normative and meta-ethical relativism. In contrast to all above, naturalist accounts typically hold that the question of normativity is bogus, and explain the apparent authority of normative claims with psychological and sociological processes of internalization and enforcement. Typically, these accounts propose a genealogy of moral values meant to undermine the special authority of moral normativity (Mackie 1977; Joyce 2001). Not all genealogies are debunking, however. Building on studies in evolutionary biology, Gibbard (1990) and Nozick (2001) argue that ethical norms and values are selected because they favor coordination to mutual benefit. While Gibbard holds that the approach supports an expressivist account of normativity and normative language, Nozick favors a more complex theory, akin to structural realism, according to which there are invariant ethical structures identifiable under admissible transformations. A plausible philosophical explanation should focus on such structures, even though there can be competing explanations of the same phenomena. Nozick's explanation relates the raise of normativity to a normativity module favored by evolution, and it is capable of defending the possibility of moral progress, which he identifies through "a multistage process whereby cooperation between distinct groups gets established" (Nozick 2001, 263). A similar conclusion about the relation between progress and agreement on a core set of practical norms is drawn by David J. Velleman, even though he emphasizes the relativist implications of the claim (Velleman 2009, 2013).

These debates further contribute to larger and deeper issues about the *objectivity of value*. On a realist standard, the objectivity of values is an issue determined by their ontological status (Harman and Thomson 1996). However, this position makes it impossible to explain how values have an impact on our life and action. If values are intrinsically motivating properties, then they are queer properties. This is the objection that J. L. Mackie moves against objectivist theories of values. It is similar

to Hume's objection that moral distinctions do not originate in reason because reason alone cannot move us to action (Hume 1739, 456–457; Schroeder 2009, 2011). Hume and Mackie assume that the authority of values and their significance to action can be understood in terms of their motivational force. Mackie concludes that the motivating power of ethical judgments shows the inadequacy of all objectivist theories of value. Hume, instead, takes this to be an argument against rationalism. Both assume that objectivity is granted by ontology. However, many philosophers suggest that the ontological conception of objectivity is useless or even dangerous in ethics. Some hold that ontological objectivity is inapplicable because there is no hope to converge on an independent reality (Williams 1985). Others argue, instead, that this is a misconception of the objectivity of value, which has to be rejected. Sensibility theorists, such as John McDowell and David Wiggins, argue that values bear an interesting resemblance with secondary qualities such as colors, insofar as the exercise of evaluative judgments is cognitive and yet concerns properties that are neither part of the causal structure of the world nor reducible to them. Our sensibilities seem to be inevitably implicated in evaluations, and this should indicate that the objectivity of evaluative standards cannot be totally independent of our distinctive sensibility. This is one way in which sensibility theories attempt to reconcile the aspiration to objectivity with the action-guiding aspiration of evaluative judgments. The outstanding question is whether they can thereby vindicate the authority or compellingness of some categories of ethical values.

Focusing on such categories, Roderick Firth and Richard Brandt propose to conceive of objectivity and normativity in terms of idealized conditions of rationality, which importantly include dispositions to respond to value (Firth 1951; Brandt 1996). By contrast, others explore a conception of objectivity centered on the principles of practical reason (Baier 1985; Toulmin 1950; von Wright). In particular, Kantians propose a practical conception of objectivity, which is based on shared reasons constructed via reasoning, under idealized conditions (Rawls 1980a, b; O'Neill 1989). It is often assumed that the practical conception of objectivity is weak and more modest than the ontological conception. In fact, the practical conception is more ambitious and demanding insofar as it includes other important dimension of values, such as their authority (Bagnoli 2013). For G. E. Moore, the objectivity and autonomy of values is warranted by the reality of irreducible normative relations (Moore 2004, 2010). On the Kantian view, instead, the autonomy of ethics is granted by the supremacy of pure practical reason. For Kant, practical reason has the unique power of producing its proper objects and should be recognized as sovereign. John Rawls has revived the Kantian conception of ethical objectivity in order to overcome an impasse in political theory, which he attributed to an inadequate conception of the standard of objectivity (Rawls 1980a, b). Neither Rawls nor Kantian philosophers are oblivious of the fact that the authority of value is often based on relations of powers. However, they think that progress can be made in building a dialogue governed by mutual respect and recognition. They also think that reasoning together is an activity constructive and productive, that is, generative of reasons that we could all share. To this extent, these philosophers believe in the reconciliatory and cooperative powers of reason.

In contrast to rationalists, relativists account for the apparent authority of moral values by invoking mechanisms of social enforcement, such as blame, and institutionalized devices to compel and coerce into compliance (Williams 1985; Blackburn 1998). On this view, the special status of moral values depends on the specific manner of their social enforcement, rather than on ontological and epistemological features of values. For Nozick, coercive enforcement is legitimate only insofar as it concerns the basic ethical norms of respect and cooperative virtues. He considers it a sign of moral progress that personal values are not a matter of social enforcement (Nozick 2001, 264–265). Gibbard distinguishes between repressive and coercive enforcement, showing that repression is always morally objectionable, while coercive enforcement may be necessary to the stability of society. Furthermore, he argues that the facts of value disagreement may be so divisive that it might be prudent to endorse principles of accommodation (Stevenson 2009).

The issue of the legitimacy of enforcement acquires center stage in liberal theories. Famously, John Rawls advocates a principle of political legitimacy according to which the exercise of power is legitimate only insofar as it respects all citizens as free and equal, hence capable to reasonably disagree about matters of values (Rawls 1993, 137). Reasonable disagreement is to be protected, rather than erased or undermined; this is a consequence of an argument that treats persons as "self-originating sources of valid claims" (Rawls 1980a, b, 582). Ontological issues appear to have crucial normative implications for law as a coercive system of norms. Should law protect values? Which values should law protect? Can values be enforced, and should they? And, most importantly, which values should be enforced? A crucial example is the case of human rights: If they are based on the value of humanity, do they bind even those who do not endorse this value? Which authority is in charge of the enforcement of human rights, in the face of a disagreement about the basis of their justification?

Both Rawls and Nozick agree that coercion and the exercise of power stand in need of rational justification. For Rawls, political power is legitimate when used in ways that all citizens can reasonably be expected to endorse. Thus, the legitimacy of a particular set of basic laws is determined by the so-called *criterion of reciprocity*, according to which citizens must reasonably believe that all citizens can reasonably accept the enforcement of a particular set of basic laws. This is an epistemic principle that governs the mutual expectations of citizens. An analogous epistemic principle governs the normative expectations of groups that recognize mutual dependence and thus coordinate for mutual benefit (Nozick 2001, ch. 5; Gibbard 1990). In a just society, citizens are in the epistemic and moral position to endorse the fundamental political arrangements freely, that is, free from manipulation, repression, and lack of adequate information. As we will see in Sect. 5, this issue is particularly relevant in the case of a pluralistic citizenry.

2 Aims and *Desiderata* for a Theory of Value

While the basic questions about values admit of different answers, there is a significant convergence about the aims and desiderata of an adequate theory of value. Theories of values are expected to explain the nature of values, specify their varieties, and account for their place in our lives. Correspondingly, the most important criterion of adequacy for a theory of value is its capacity to offer good philosophical explanations of why and how we should value things, objects, and activities. Call this the *explanatory capacity* requirement. Further requirements apply insofar as the basic aim of the theory is to produce convincing explanations of phenomena concerning our evaluative practices. Arguably, a good explanation is coherent. *Coherence* is a logical requirement, which applies especially if the theory of value is systematic and normative, but it seems also important for those theories of value based on a rational justification of value. In view of debates about the nature and ontological status of values, there is a consensus about the criterion of *ontological parsimony*, according to which one should not introduce more entities than they are required to explain the phenomena. However, philosophers disagree about the sort of properties that perspicuous explanations should admit in order to make sense of valuing practices and discourse. Theories of value are also expected to account for whether and how we know values, or else explain why the quest for knowledge of this sort is misplaced. Concerning the relation between facts and values, an adequate value theory seems superior to others when it explains why some values or clusters of values are indispensable to explain a range of facts. In other words, a value theory places values in the best explanation of what we experience in the world. According to some, values do not figure in the best explanation of what we experience in our evaluative practices; hence, they are dispensable entities (Harman 1977, 2000; Harman and Thompson 1996). According to others, instead, values have a place in the best explanation of how things stand (Sturgeon 1986, 1998). This position is supported by various versions of the *open-question argument*, originally devised by G. E. Moore. Moore argued that for any definition of good in terms of other properties, it is an intelligible question to ask whether such a property is good. This means that the meaning of the concept of good is not exhausted by its analysis. The implication is that matters of values cannot be explained away in terms of matters of (natural) facts, and that any reduction of value in terms of other properties leaves the crucial question unanswered (Moore 2004, 44). This argument is directed to any attempt at reducing value to other properties, either metaphysical or natural. Moore also argued that there is a naturalistic fallacy involved in such a reduction, but many critics have pointed out that there is not (Frankena 1939). Naturalists claim that the only properties necessary belong to a naturalist ontology, while realists claim that we should admit also of moral and evaluative properties, which are not identical with or reducible to natural properties. Subtler disagreements concern the very definition of natural and non-natural properties, and the forms and status of naturalistic reduction.

Thomas Scanlon offers a reductivist analysis of the concept of good in terms of reasons, which does not seem vulnerable to the open-question argument. His

"buck-passing" analysis avoids indicating which properties make and object good. The analysis is supposed to show that once the evaluator has identified the reasons she values something as good, there is no further work for the concept of value to do. This is the claim of redundancy of value (Scanlon 1998, 97). Furthermore, the analysis of value in terms of reasons demystifies evaluative concepts and dispels the aura of gravity and inexplicable compellingness that surrounds terms such as good (Scanlon 1998, 98).

Disagreements about the possibility of reducing value to other concepts are relevant to determine the perspicuity of explanations. According to a general criterion of *descriptive plausibility*, a theory should explain why values appear as they do. For realists, the criterion requires that we consider values as part of the fabric of the world and take evaluations at face value, as assertions about values. Instead, others hold that a good explanation of value should fit our ordinary understanding and practices of value, and propose a requirement of *congruence* with the subjective experience of valuing we have as agents and citizens. On this view, it would be implausible to offer an account of values that systematically discounts and undermines the subjective experience of evaluators. By contrast, error theorists adopt a weaker criterion that allows an adequate theory to explain away the appearances of our evaluative practices as due to some systematic illusion or error (Mackie 1977). Likewise, projectivism holds that evaluative judgments formally behave like assertions about what is the case, but they are in fact the result of subjective projections or patterns of objectification (Blackburn 1984). Through such patterns of objectifications, values have gained a relatively solid ontological status, even though they originate in our mind and social practices, rather than being features of the world. These weaker criteria of descriptive plausibility seem problematic, however. On the one hand, it is unclear whether their semantic apparatus suffices to warrant our claim to truth. On the other hand, their analysis of value judgments as involving a systematic error or projection seems ultimately self-defeating: It corrodes moral authority and demands such a radical revision of ordinary evaluative practices that they become unsustainable as ordinarily known. An adequate explanation must not be self-effacing; that is, it should be one that does not undermine or undercut the importance attributed to moral values in ordinary thinking. I will call this the requirement of *reflective stability*, and others call it *transparency* (Korsgaard 1996, ch. 1).

In addition to criteria that measure the explanatory capacity of a theory of value, there are others that are meant to identify its *normative capacity*, that is, the capacity to guide action, attitude, and belief. It is a criterion of adequacy of value theory that it accounts for the relation between *values and rational requirements*. This criterion is interpreted in different ways, according to the various theories of rational authority. Consequentialists reduce all issues of deontic relations to value. The contrary view reduces value to deontic or normative relations, for instance the relation of reasons. This seems to be the case of the so-called buck-passing account (Scanlon 1998, 97), which reduces claims about the value of an object to the reason-providing properties of the object. The concept of value can thus be analyzed in terms of reasons and the properties of objects that provide them for evaluators.

In relation to the normative capacity to guide action, some philosophers hold that a theory of value should also account for the fact that at least some values are motivating, other things being equal. As mentioned above, Humeans hold that the authority of value amounts to its motivational and conative force. Endorsing the Humean perspective, some further argue that adopting a value commits to being motivated and bringing it into the world. For some, such a commitment is intrinsic to the very idea of adopting a value, so that when we endorse a value judgment we are thereby motivated to act accordingly, absent any incapacitation, interference, or impediment. On this view, the relation between value and motivation is thought to be internal and conceptual. For others, instead, there is no such a relation. When values motivate, it is because of an intervening external factor, such as a desire or an interest, whose independent force motivates action in conformity to value judgment. If Spartacus promotes equality, it is because he wants to promote equality, not merely because he thinks that equality is valuable. This desire attaches to and accompanies the evaluation and works as a trigger of action, a force external to the evaluator's practical reasoning and judgment. By contrast, Kantian philosophers such as Thomas Nagel defend the view that practical principles themselves generate desires (Nagel 1979). For instance, the desire to promote equality is generated directly by the adoption of the principle that prescribes equality. Furthermore, Kantians do not think that the normative capacity is exhausted by the capacity to motivate action. On the contrary, they hold that the key notion that articulates the normative capacity of value is that of rational requirement. Their view is that action motivated by the very idea of duty is of special value. By contrast, actions that conform to duty, but are motivated by self-interest or inclinations, do not have moral value. Such views link theories of value to moral philosophical psychology; hence, the debates about standards of objectivity are placed at the juncture among different disciplines. A further methodological issue is whether such disciplines are thoroughly empirical or may accept a priori arguments, concerning for instance the conditions of possibility of valuable action or the form of a moral agency (Anscombe 1957, 1958; Murdoch 2013; Thompson 2008).

3 Some Substantive Questions About Value

In addition to meta-ethical questions about the reality of values, the role of evaluative practices, and the logical grammar of evaluative judgments, there are substantive questions about value. What things have value, and how so? There are many things that seem good and a variety of ways in which they are good. Pleasure, knowledge, beauty, and personal relations are natural candidates and often thought of as indispensable ingredients for a flourishing life. Some theories take this variety seriously and hold that there are many different values and many ways in which things are good. These are pluralist theories, and they vary according to which goods they include and how they think such goods are related, if they are related at all. By contrast, monistic

theories recognize only one kind of value. For instance, hedonism takes pleasure to be the only value, and utilitarianism takes utility to be the only value.

Value is a generic term that is attached to several varieties of items, which can be grouped into three basic categories of values. Values can be attached to (i) actions, policies, and institutions; (ii) objects, plans, projects, lives, prospects, events, states of affairs, outcomes, consequences, and effects; and (iii) to character, character traits, dispositions, intentions, and actions understood as bearing the marks of agency. The first category is generally governed by concepts such as right and wrong. The second category is understood as the sort of goodness that identifies something as worthy of choice. A fair taxation is good because it brings about good effects, promotes equity, and favors the least advantaged. The third category is often qualified as moral, in that character and intentions are traditionally objects of moral evaluation and assessment. There are important disagreements also regarding how we value things belonging to these categories. For instance, ethical theories take items in the third category as special sorts of values, but they differ in accounting for what makes them especially valuable.

When focusing on the act of valuing, it is useful to distinguish three kinds of value judgments. First, evaluators value an object because of the specific features it has as an object of a certain kind. A diamond is good because of its particular luminosity, purity, and saturation. Second, we may value something as a good thing; for instance, when we say that peace is a good thing to live or die for, or to bring about. Third, we may regard some things to be good absolutely, in that the world is better because of their existence; beauty, love, and friendship may be proposed as instances of this kind (Kolodny 2003).

When we say that some things are good, one may intelligibly ask "Good for what?" This question does not seem always pertinent, because it is often the case that the answer is implicitly contained in the description of the object that is said to be good. For instance, it may be superfluous to ask what a knife is good for, insofar it is obvious what the function of a knife is and also what properties make it suitable to fulfill its function. In other cases, instead, how to identify the relevant function is more problematic. What is the function of a human being? Is there any one function that fully defines what humans are? What is the function of government? What functions do moral virtues help us realize? Inspired by Aristotle, many seem to think that the disanalogy is only apparent and that all evaluative judgments admit of a similar analysis: They are relative to a function, even though reference to such function is often implicit. In support of this view, some have argued that the term good is a predicate modifier, rather than a predicate (Aristotle 1984; Geach 1960).

The disagreement about the function of the concept of good is particularly relevant in ethics as it bears normative and deontic implications. When the question "for what?" applies pertinently, then the value at stake is instrumental. For instance, practicing piano everyday is good for learning to play piano and drinking water is good for health, money is good for buying things we need. The deontic implication of these propositions is that money is good only to the extent that and insofar as it is a means to get some other goods, e.g., an object, a title, or a status. Many objects and activities are instrumentally good, but are all goods instrumental values? Some

philosophers hold that goodness is a property, which can be predicated also *simpliciter*. In this case, the value at stake is intrinsic. A typical example of intrinsic value is the value of persons, even though this claim is not uncontested. Utilitarians and Aristotelians identify happiness as the overarching value, but they define it in different ways. For pluralists, happiness is a complex condition to which many different goods contribute. Some of these goods are objects, activities, or personal relations.

The difference between instrumental and non-instrumental values can be explained in terms of perspectives. What is good for *F* is good for her own perspective. Being good *simpliciter* means being good from the point of view of the universe, hence from no specific perspective at all, i.e., good "from nowhere," that is (Nagel 1986). Agglomerative theories, instead, hold that goodness *simpliciter* amounts to the sum of all perspectival goods. For instance, utilitarianism is committed to this view. Egoism typically holds that what is good *simpliciter* is defined by what is good for the evaluator. A distinct but related question is whether the bearer of value is always a state of affairs. Some are inclined to think that if good concerns states of affairs it is always attributive, and then goodness *simpliciter* can be analyzed in terms of what it is good for. By contrast, others hold that it is a mistake to conceive of the constitutive aspect of valuing in terms of bringing about a state of affairs.

On some theories, the distinction between instrumental and final goods collapses on, or is equated to, another distinction between intrinsic and extrinsic goodness. Instrumental goods are also held to be extrinsic goods, and final goods intrinsic goods. However, the distinction between intrinsic and extrinsic goodness concerns the sources of values, while the distinction between final and instrumental goods concerns the way we attach value to items. To bear in mind this distinction allows us to appreciate that there are things valued as extrinsically good, and yet as final ends. For instance, painting or horseback riding might be final ends because of the interest we take in them. Separating these distinctions allows us to distinguish two ways to treat this case. Some realists hold that the goodness of final ends is intrinsic, absolute, and independent of interests and desires (Moore 2010). By contrast, Kantians allow for extrinsically valuable ends that are valuable because people take an interest in them (Korsgaard 1983). Such ends are things we want for there own sake, but whose justification resides in something else, e.g., in the fact that we want them. They are not unconditionally good, since their being good is relative to us, but they are nonetheless rationally justified on the basis of their desirability.

4 Theories of Value

Traditionally, the theory of value represents one of the two main branches of ethical theory, along with the theory of right. While the theory of right tells us what we ought to do, the theory of value specifies to what end we should act, what states of affairs are good or bad, hence the distinction between virtue and vice. We can distinguish five main theories. *Hedonism* is the view that only pleasure is intrinsically good, and

only pain intrinsically bad. This view admits of many different formulations, which vary in the definition of pleasure. According to some, pleasure is a sensation which is felt as pleasant. According to others, pleasure is a sensation that people want to have because of its subjective quality or sensation. The latter formulation seems to include a richer view of value that the term pleasure does not capture. A key issue for hedonism is measurement, since in order to justify determinate evaluative judgments, one must clarify how subjective sensations can be compared and ranked. For Jeremy Bentham and Henry Sidgwick, pleasure and pain are symmetrical, but others hold that pain is a greater evil than the absence of pleasure, and attribute greater ethical significance to intense pain.

While the view that pleasure is good and pain is bad is very plausible, the hedonist claim that pleasure is the only good is highly controversial. A crucial argument against this view is produced by Robert Nozick, and it involves the thought experiment of the pleasure machine (Nozick 1974, 43). This is a fictional case in which we consider whether to be plugged into a machine that gives us pleasure by interacting with our brain. If hedonism were right, then we all have decisive reason to plug in. By contrast, Nozick argues that there are three reasons not to plug in, all referring to how we care about knowledge. First, what matters to us as agents is not simply to have the experiences associated with actions; rather, agents want to be the persons who do such actions. Second, agents care about being persons, rather than a mass floating in a tank capable of feeling pleasure. Third, agents care about experiencing actual reality rather than a simulation. The pleasure machine simulates the experience of pleasure without the subject experiencing any actual reality. It matters to us that we feel pleasure through experiences, without losing grip of reality.

Desire theories argue that pleasure is too restrictive a notion, insofar as one might desire goods and experiences that are not themselves pleasurable, as Nozick's pleasure machine thought experiment shows (Nozick 1974, 43). In Nozick's case, knowledge of how things stand and acquaintance with reality is a more desirable experience than the comforting and pleasurable sensations grounded on false belief. This shows that there are things that are important and matter to us but not because they are pleasurable. As for hedonism, there are different formulations of desire theories. One is that good is a state of affairs where the agent obtains the good he desires. Others commit to more normative claims about what constitutes the well-being for a person, which may differ from what one actually desires in particular circumstances. It is controversial whether this kind of theory avoids Nozick's objection, insofar as it seems committed to say that whenever people desire to be plugged into the pleasure machine, this is good for them. It seems that to avoid Nozick's objection, one needs to offer stronger grounds.

Perfectionism identifies a plurality of goods and typically regards knowledge as the highest value. This theory has been defended in various forms, from Plato and Aristotle, to Aquinas, Hegel, Nietzsche, and G. E. Moore in the analytic tradition. Some perfectionist accounts emphasize the plurality of goods as a ineliminable feature of the theory; others recognize the plurality of goods, but they also insist on the importance of giving a unitary account in the explanation of goods, and also, and more importantly, to organize the plurality of concrete goods under a more general

and abstract characterization or formal structure (Hurka 1993, chs. 8–10). Perfectionism about the good is typically combined with an account of the virtues that make such goods achievable and realizable.

Since the goods are many, and virtue is what makes us fit for the good, the question arises as to how to harmonize the virtues. Perfectionists in the early twentieth century, such as G. E. Moore and W. D. Ross, do not offer a unified theory of the virtues, but rather insist on the distinction between things that are good and dispositions of the mind and of characters that qualify as virtues. However, these perfectionists were sensitive to the explanatory question raised above and tried to account for a common ground that explains both why some things are good and why some character traits and dispositions of the mind are virtuous (Moore 2004, 204, 208–211; Ross 1930, 134–160). Instead, more recent perfectionists argue against the view that a unified explanatory account is needed (Adams 2006, 31). More ambitious theories aim to offer a unified theory of value and virtue. The first prominent example of this case is Aristotle. On the Aristotelian account, the most accomplished human life consists of a complex hierarchy of material goods, excellences of character, and excellences of the mind, which all work together toward the realization of the human being as a rational and political animal. Among all activities that are distinctive and typical of human beings, there are cooperative activities among friends.

5 Value Disagreement

People disagree about what to value, why, and how. Some love knowledge, and others search for pleasure. Some think knowledge is valuable because it helps us survive and others claim it is good in itself. Further disagreements concern the normative implications of valuing objects such as knowledge, pleasure, or people. What kinds of normative commitments and attitudes do valuing require? If one values knowledge, one is committed to promoting knowledge. If one values pleasure, one has a normative reason to engage in activities that are of one's satisfaction. If one values persons, does one have the same sort of normative reasons as in the previous cases? It seems that valuing persons requires different normative attitudes than those associated with realizing states or affairs or bringing about consequences. This is apparent in the case of conflicts of values where the life of others is at stake and there is no policy that relieves the agent from guilt.

Some argue that value disagreements are more widespread and more pervasive in ethics than they are in science (Harman 1977; Williams 1985). They remark that even when there is an agreement about basic facts of the matter, a value disagreement may persist. For instance, conservatives and liberals may handle the very same data about stem cell research and yet disagree about the moral and political permissibility of it. How to understand such disagreement and what conclusions to derive from the apparent facts with which we do disagree is an open question. However, the brute fact of value disagreement has supported two meta-ethical positions: nihilism about the existence of value and skepticism about the possibility of moral knowledge.

These two positions are often conjoined, but they support different claims about the nature of value. Nihilism is the view that there are no values and it does entail that there cannot be knowledge of values, where this is understood to be knowledge of peculiar objects. However, moral knowledge may be construed differently, as a kind of practical knowledge that originates in a special relation one entertains to oneself as a practical subject. In this distinctive sense, moral knowledge does not require any moral ontology, and thus, it is not in contrast to nihilism. Furthermore, one may hold that values exist, but lie beyond our reach. This would be a case in which skepticism about moral knowledge presupposes realism about value and makes the condition of possibility of knowledge so distant from our standards that they are never met, and thus, knowledge is never obtained. This is a logically consistent position, but of little practical import. Normally, however, nihilists tend to be skeptical also about the possibility of moral knowledge.

Whether widespread and persistent disagreement about value demonstrates nihilism and supports skepticism is a controversial and complex matter, which we can begin to address by considering the sort of disagreement that would challenge moral truths and the existence of values. A first consideration is that disagreements about values can be identified and discerned against the background of shared practices. For instance, liberals and libertarians strongly disagree about the sort of cases in which it is legitimate to use coercive instruments to redistribute wealth such as taxation, but this disagreement is intelligible only against the background of shared practices informed by the principle of political legitimacy, which requires one to justify the deployment of coercive enforcement. Second, interesting disagreements often concern specific claims and are intelligible against the background of general principles. For instance, different traditions have delivered very peculiar catalogs of the virtues, the excellences of character. Philosophers as diverse as Aristotle, Hume, and Kant disagree about the significance of wit, or the morality of magnificence, but they would broadly agree that inflicting unbearable suffering for fun is morally objectionable or that unqualified pain is a disvalue (MacIntyre 1988, 2007, 179–181; Scanlon 1995). This seems to suggest that there are some very basic and general norms that are invariant, while more specific moral norms vary across societies. Naturalists explain this fact on the hypothesis that the core ethical norms favor coordination for mutual benefit and different societies might find different equilibria (Gibbard 1990; Nozick 2001, ch. 5).

It is noteworthy that there is an asymmetry about how the facts of value disagreement and agreement are used in the debate about the ontological and epistemological status of value. Persistent value disagreement in the face of factual agreement is thought to be evidence for skepticism and nihilism; but the presence of large areas of agreement and concordance is not typically used to show that, at least on such areas moral knowledge can be obtained (Nagel 1979; Parfit 1984, 452–453). Third, some philosophers insist on vagueness as a possible epistemic source of disagreements. On this analysis, our value disagreements depend on the fact that we do not know what is the correct position on value matters, even though there is one correct answer in each case (Brink 1984; Boyd 1988; Shafer-Landau 1994). There are interesting cases of value disagreement that rest on factual ignorance, irrational ignorance,

self-deception, or implicit bias, and thus can be fruitfully explained otherwise, without endorsing nihilism about values. For instance, two colleagues might disagree about the opportunity to increment diversity in a department, not because they disagree about the value of diversity or the proper means to implement it, but because they do not have access to the same facts about the presence of minorities in positions of powers or holding offices. Furthermore, another colleague might disagree about the same issue, not because she does not have access to the relevant facts of the matter, but because she self-deceptively resists the evidence that the department is not as diverse as she thinks it should be, discounting the facts that would support a different belief about the state of the department and its recruitment policy (Shafer-Landau 1994). These are interesting analyses because they uncover some complexities in the case of value disagreements, whose roots are not always identifiable as either factual or evaluative. Evaluators often disagree not only on facts per se, but also on their relevance and importance, on the nature and strength of evidence provided, and on the question who bears the burden of persuasion.

There are different perspectives about the significance of genuine value disagreements. The presence of irresolvable value conflicts is generally taken to undermine the aspiration to objectivity of ethics. For instance, Bernard Williams argues that the pervasiveness and untreatable character of conflicts of values shows that the aspiration to ethical objectivity is not a matter of logic and cannot be understood in terms of convergence on an independent moral reality (Williams 1985). There is a pressure toward resolving disagreements, but this is to be understood as a psychological and sociological need, functional to peacefully living together. Conflicts do not reveal any pathology of practical thought, but exhibit the richness and varieties of values. Thomas Nagel shares this diagnosis and further concludes that what we call ethics is not a homogeneous field, but a complex and heterogeneous cluster of claims, which does not admit of a systematic treatment. Nonetheless, there is a standing request for objective standards of correctness for evaluations, which in some domains may allow for weighting reasons (Nagel 1979, 180). A more optimistic view emerges as we distinguish levels and layers of disagreements. According to Scanlon (1995), this approach allows us to recognize broad areas of agreement. This is not to deny that there are radical disagreements in values, but they are not as large as the skeptics believe. When we clarify how we differ in values, we find large areas of consensus and we can further consider whether aiming to erase all sorts of value disagreement is morally appropriate.

How to decide on the latter issue, how to treat radical differences? Following Rawls, Scanlon holds that this is a matter for political philosophy, rather than for epistemology or ontology. The latter position has been argued extensively in political philosophy, especially as an argument for political liberalism. According to Rawls, reasonable citizens accept the burdens of judgment, and this is why they are inclined to tolerate disagreements, when this does not undermine or violate human rights. Reasonable citizens recognize that it is difficult, if not impossible, to settle disagreements of values, which are the root of political, religious, and moral disagreements. They do not have to deny that there ultimately are definitive answers about truthful values, but they have to admit that there are such serious epistemic limitations that

tolerance of disagreement is the most reasonable option open to them. To coercively enforce or impose by repression, a value that is not shared would be a violation of freedom. In dealing with untreatable disagreements, Rawls addresses the problem of legitimacy within a liberal society, and his argumentation is based on the claim that the citizens of a liberal society would share not only the political institutions of a constitutional regime, but also the public traditions of their interpretation, as well as historic texts and documents that are common knowledge (Rawls 1993, 13–14). That is, they share a public political culture. In addition, they conceive of justification as a thoroughly public affair, that is, as a normative practice addressed to other citizens within the framework of a liberal society (Rawls 1993, 101). Apart from its relevance in debates about political legitimacy, this view is representative of a distinctive position in the epistemology of value disagreement. Rawls points out that the practice of normative justification governed by mutual respect and recognition has the potential of developing a free and reasoned agreement in judgment. Rawls' main concern in dealing with value disagreement is to legitimately ground the stability of pluralistic societies. However, the possibility of developing a basis of agreement in judgment is one of the foci of the debate about value pluralism. Agreement on some core values, such as respect of the dignity of others, is something that many identify as a sign of moral progress. But protecting value disagreement is also largely considered a sign of moral progress. Is it possible to defend these positions solely on political grounds, without taking sides about the nature of value?

6 Pluralism and Incommensurability

The fact of value disagreements requires a plausible explanation, and the theories of value offer competing ones. Monist theories of value reduce such disagreements to ignorance, epistemic vagueness, and bad reasoning, hence implying that there cannot be genuine disagreement about value. Pluralist theories, instead, take seriously the facts of value disagreement and try to account for plurality. In fact, how to account for plurality or the lack thereof is a challenge for both monism and pluralism. By discounting value disagreements, monist theories commit themselves to offer a plausible explanation of the sources of spurious disagreement, while pluralist theories commit themselves to explain how we can choose and behave rationally in contexts marked by pluralism. Neither task proves to be easy to accomplish. The strategies differ in relation to what we take value pluralism to be. Clearly, no theory can deny that subjects recognize different values and goods. However, pluralist theories take seriously the phenomenon of value disagreement, this plurality as irreducible, and account for such irreducibility on the basis of the claim that plural values are incommensurable. By contrast, monist theories hold that differences among values are superficial and can be explained away. This dispute over the sources and nature of value has significant normative implications, especially regarding the case of conflicts of values. To account for such implications, it is useful to distinguish different interpretations of value incommensurability.

In its stricter formulation, incommensurability is the claim that there is no common unit of measurement of value (Wiggins 1987; Stocker 1989, 175ff.; Chang 1997; Finnis 1980, 113ff; Wong 1989, 1992). Incommensurability comes in two varieties. *Weak incommensurability* holds that there is no single cardinal scale by which every value can be measured. *Strong incommensurability* holds that between any two particular values, there is no single unit by which they can be measured (See Wiggins 1987, 259; cf. Richardson 1995, 104–105). Weak incommensurability is often combined with the view that some values, such as human life and rights, have a special axiological and normative status (Anderson 1993, chs. 7–9; Lukes 1997). Supporters of this view remark that it is congruent with common valuing practices about the sacredness of life and treat some goods as radically different from commodities. Among such practices, there are procedures for establishing legal and moral constraints on the legitimacy of treating human life, personal relations, relations to non-human animals, and to the environment, along with marketable goods. For instance, this is the foundation of legal and moral arguments against slavery, prostitution, exploitation, and the trafficking of human organs.

This view carries two important implications. First, the claim that values do not differ in kind and are measurable by a single unit of value facilitates the decisions of public policies, but it undermines common evaluative practices and sensibilities, and thus, it requires a process of revision and adjustment. Second, the recognition of the irreducible varieties of values seems to require the development of an appropriate sensibility, receptiveness, and emotions (Stocker 1998). As Elizabeth Anderson remarks, different values demand different evaluative attitudes and responses from evaluators (Anderson 1993). To ensure that the recognition of the plurality of values be effective from a normative point of view, evaluators should be endowed with psychological resources apt to appropriately respond to this axiological variety (Bagnoli 2011; Deigh 2008; Slote 2010). This axiological and psychological complexity certainly complicates matters, and presumably it severely constrains the legitimacy of value transactions, but it also ensures that our lives are rich and nuanced. Vice versa, the claim that all values are ultimately the same and measured by a single unit of value simplifies our economic transactions, but it also impoverishes and flattens our emotional lives (Nussbaum 1990, 116–120; Anderson 1993).

An interesting section of this debate deals with a more practical dimension of incommensurability as it invests the life of any entity entitled to assess and capable of evaluating. For instance, Isaiah Berlin defines incommensurability as a claim about abstract values that cannot be jointly realized in the world (1969, 49–50, 53–54). On his reading, incommensurability amounts to *incompatibility*, in a given context. Insofar as it originates in concrete situations because of contingent features of the context of choice, it does not raise any logical issue about the incoherence of value system. Ronald Dworkin talks of incommensurability of specific instances of values of which one cannot say that they are always as good as, or better than, instances of other values. On this interpretation, incommensurability generates a phenomenon called *trumping, which happens when a certain sort of considerations overrides any other sort of consideration* (Dworkin 1977, xi). Trumping assumes incommensurability because it is not merely the case that some considerations are stronger than

others. On the contrary, trumping occurs when a category of considerations overrides other considerations of a totally different sort. In particular, Dworkin holds that rights trump other considerations because they have a special normative force, which insulate them from trade-offs (Dworkin 1977). For instance, advocating the trumping power of the right to free expression, John Stuart Mill writes that one cannot be silenced by the majority as much as the majority cannot be silenced by one (Mill 1988, 20). A similar view, called *discontinuity* or *threshold lexical superiority*, is designed to capture cases in which some threshold amount of one value trumps any amount of the other value (Griffin 1986, 85).

These formulations of incommensurability serve well deontological ethical theories, where the categories of duty and right are supposed to be separated from the category of utility. For some philosophers, incommensurability is a *constitutive* feature of values such as respect for persons or friendship (Raz 1986, 345–357).

Kant distinguishes between value and price (Kant 1997, 4.432). Things have a price and thus are inter-substitutable and mutually fungible. These features warrant the commensurability of commodities, which is the basis of market relations. By contrast, dignity does not admit of measurement, and this blocks any form of compensation in kind when different persons are at stake. The opposite view is typically associated with monist utilitarianism, such as Jeremy Bentham's theory, which takes pleasure to be the only value, in terms of which all other values could be measured. More contemporary versions of utilitarianism take informed preference as the basis of such assessments, but they all adopt commensurability (Hare 1981; Harsany 1974). While commensurability is typically associated with consequentialism in its utilitarian variants, it is arguable that consequentialism may recognize at least weak incommensurability. For instance, some philosophers argue that while it is true that there is no rate of substitution among values and thus no general maximizing principle that can guide action, there is a general duty to bring about the best consequences (Finnis 1980, 113; Stocker 1998) . The content of duty is not uniform across contexts because the definition of what counts as the "best consequence" is relative to incommensurable values.

Strictly speaking, comparisons and rankings do not require commensurability. Therefore, the philosophically significant problem is comparability. Unfortunately, the claim of strict incommensurability (i.e., the lack of a cardinal unit by which values can be measured) is not always neatly distinguished from various failures of "incomparability" (i.e., lack of ranking relatively to a covering value). Considerations about how to rank values often merge with considerations about the limited cognitive and practical capacities of evaluators, which result in imprecision and indeterminacy. This idea turns on the claim that comparability is a matter of precise cardinal comparison. Derek Parfit holds that there are cases of imprecise cardinal comparisons (Parfit 1984, 431ff.). When incomparables are really indeterminately comparable, that is, it is neither true nor false that they stand in a positive value relation or "rough comparability" (Griffin 1986, 80–81, 96) or "vagueness" in comparison (Broome 1997). A general view of value that might explain the last three phenomena is that values are not determinate quantities but are metaphysically indeterminate (Chang 2002, 143–145).

Most philosophers in this debate hold that there are three positive value relations of comparison: "better than," "worse than," and "equally good." By contrast, Ruth Chang suggests that there is a fourth basic value relation, called "parity," which indicates a relation between two objects different than equality. Whether parity is really a distinct fourth value relation, not reducible to equality, is a debatable claim, but there are two important considerations in its support. First, parity seems indistinguishable from equality only on the presumption that values should be modeled on the relations among real numbers, which is a contestable claim (Chang 2002). Second, there are comparisons that do not seem to fit the standard tripartition outlined above. For instance, suppose an expert consultant is required to assess a policy regulating promotion in a research laboratory and concludes that: "*x* is a better policy than *z* because it is meritocratic." The consultant does not thereby imply that *x* is a better policy than *z*, absolutely. Her comparison is constrained by the context, which is limited to evaluating research and, therefore, assumes some values that are relevant in research assessment exercises. Were she required to compare the two policies in relation to another context, e.g., a public geriatric hospital, the consultant's judgment would likely be very different because it would be informed by other values, e.g., the values of health care. It would be grotesque to consider meritocracy as a dominant value in regulating patients' admission. The example shows that comparisons are made according to a range of considerations governed by substantive values, such as merit or fairness. Formally, then, comparability is a three-place relation: For any value *x* and *y*, *x* is comparable with *y* with respect to *V*, "a covering consideration" (Chang 1997). Also conversely, incomparability is a three-place relation: *x* is incomparable with *y* with respect to *V*, where *V* is a covering consideration. On this interpretation, then, incomparability does not amount to non-comparability, which holds when the formal conditions required for comparing values are not met.

Chang introduces an important complication in the relation among values, but does not challenge the basic assumption that deliberation and rational decision require comparing or commensurate options. A profound question is whether commensurability is a morally appropriate requisite for any transactions about value (Scanlon 1991). A radical view is that at least when some important values are at stake, comparability is a misplaced expectation and inappropriate method (Raz 1986, 322; Lukes 1997, 185–186). According to Steven Lukes, this method encourages an impoverished and sometimes even corrupted conception of value. To ask for comparison is a moral mistake, which aptly attracts blame. Conversely, refusal to compare shows that the evaluator correctly comprehends and understands the practice of value (Lukes 1997, 185–186; Raz 1986, 345–357). This radical view makes clear that the shared assumption about commensurability as a requirement for rational choice commits to commodification, that is, the claim that all sorts of values are like commodities.

The advantage of this approach is that practical reasoning about what to do becomes a form of calculation. An important example of the role of cardinality is the discussion of well-being, which spreads from ethics to welfarist economy. For instance, John Broome (1991) argues that how people's states of well-being (i.e., how well off they are) should be aggregated in order to determine the value of an overall distribution of well-being. According to "the interpersonal addition theorem,"

for a single group of people, one distribution of well-being is better than another if, and only if, the weighted total of the well-being of its members is greater. The same reasoning applies when we compare the values of distributions of well-being for different groups of the same size and then groups of different sizes; and, similarly, when we compare the states of well-being of a single person at different times (i.e., states of "temporal well-being"), in order to determine her overall, lifetime level of well-being. Broome concludes that the lifetime well-being is the sum of the values of all of one's states of temporal well-being, and he proposes that the value of a distribution of well-being for any population is the sum of the amounts by which the lifetime well-being of each person exceeds the neutral level for adding a life (Sidgwick 1907). Broome's demonstration requires that well-being can be measured cardinally. A problem typical of this approach is that it seems counterintuitive and unfair. For instance, it generates what Parfit names the repugnant conclusion, that is, that given a population of any size in which everyone enjoys an extremely high level of well-being, it would be better to have a much larger population of people, all of whose lives are barely worth living. Further qualifications about what it takes to have a life worth living may lessen the problem of unfairness, but do not solve it.

Are there compelling reasons to accept the claim about commensurability? Elizabeth Anderson argues that there are not. Commensurability is dispensable because it is not really useful. It does not make any sense to call objects good when they bear no relation with agents (Anderson 1997, 91; Slote 1989). Anderson defends a pragmatist theory according to which judgments of value are constructions of practical reason that guide practical reasoning. This is to say that judgments of value make sense within a practical domain, when we consider what to do. She thinks this conception commits to the view that we value things only insofar as they relate to us in some significant way, and also that we can justify or value judgments only pointing to practical functions (Anderson 1997, 91). A further implication of the pragmatist conception is that rational deliberation is never only about means, but always also about the ends. That is, values are always implicated when reasoning about what to do. This view has the merit of showing that discussions about the nature of value are strongly connected to rational choice and to issues such as the integrity and practical identity of agents.

By contrast, there are attempts to show that pluralism is an illusory phenomenon, which can be reduced to monism, without any loss of descriptive plausibility. Attempts of this kind typically distinguish between the subjective experience of values and the real ontological stance of value, e.g., a projectivist story about how values become part of the fabric of the world (Mackie 1977; Blackburn 1984, 1985). However, it is arguable that this reduction generates loss of descriptive plausibility. It is preferable a theory of value whose full-fledged epistemological and ontological story does not routinely and systematically discount subjective experience as illusory, but it is congruent with it.

On the basis of this argument, pluralists argue that monism is false to facts, while pluralism exhibits a high explanatory capacity. We have seen how it can explain the language of rights and the special value commonly attributed to personal relations. Furthermore, the pluralist claim helps us explain several predicaments of practical

rationality. This is an important reason for placing incommensurability at the center of debates about the powers and limitations of practical reason. For instance, value incommensurability may be as one of the possible sources of *akrasia.* We generally take this case to be such that the agents do not conform to duty even when they know what they have to do. However, one other account of the phenomenon is that their reasoning does not fully determine what to do, but prescribes different and incompatible lines of action. The break in the line from the reason for action to action is not at the level of motivation, but it is situated earlier, in the contrasting values that inform the starting point of practical reasoning. The claim that the plurality of values blocks inter-substitutability helps explain why in the presence of different valuable ends, agents may respond with *akratic* behavior (Wiggins 1987, 239; Stocker 1989, 230ff.). It is debatable whether such phenomena are merely subjective illusions due to the cognitive and practical limitations of human psychology or instead depend on the ontological features of the value domain. The capacity to explain the phenomenology of valuing is certainly an asset of value pluralism. It may be objected that this is a consequence of its incapacity to guide choice.

7 Values and Rational Choice

The main significance of incomparability is that it threatens the possibility of rational choice. If two alternatives for choice are incomparable with respect to the values that matter in the choice between them, then, it is widely believed, there can be no rationally justified choice between them. Some admit of "existential plumping" for one alternative over the other (Chang 1997, 11; Broome 2000, 33–34). Incommensurability covers a wide range of phenomena where in case of conflicts, there is no principled way to rank the values at stake and no independent value that may be invoked as the umpire (Williams 1981). This view has three normative implications. First, from a normative point of view, it implies that value conflicts generate moral dilemmas where obligations clash, or practical dilemmas where reasons for action clash, and there is no resolution based on reason. In such dilemmatic contexts, decision always involves a loss in value. Second, this loss justifies and is congruent with the emotional experience of regret or guilt. Third, the phenomenology of choice is marked by attitudes and emotions that count as moral residue, even when there are attempts to trade off values, and within practices where compromise and compensations are legitimate.

Because of these profound implications for choice, some hold that the incommensurability of values undermines ethical theory as a theoretical and practical enterprise because it makes it impossible to provide a coherent and efficacious method of practical reasoning to help the agent determine what to do (Hare 1981). Value pluralism adds levels of complexity to monism, but it seems to undermine the rational basis for choice. A monistic account of value facilitates rational choice, but it oversimplifies the experience of valuing and it correspondingly impoverishes our evaluative life, seeming false to our experience as evaluators. The dispute between monists and

pluralists concerning rational choice is articulated around the two meta-theoretical desiderata: normative determinacy, which is the alleged prerogative of monism, and descriptive plausibility and congruence, which are central for pluralism.

However, there are two orders of considerations for subverting either of these conclusions. The first order of considerations concerns the possibility of ranking incommensurable values so as to justify rational choice. Pluralists do not deny the possibility of ranking values, but argue that such orderings are not complete and admit of partial, dominant, and vague orderings (Sen and Williams 1982, 17). As mentioned in the previous section, most debates are based on the assumption that the lack of a cardinal scale for values prevents the comparison of instances of values. This assumption is false because the lack of a cardinal scale of measure does not entail incomparability (Bagnoli 2000, 2006; Chang 1997).

The second order of considerations for denying that pluralism undermines rational choice concerns the form of practical reasoning applicable in pluralist contexts. Pluralists have devised several strategies to rationally justify choice in pluralistic contexts, and all deploy rational deliberation. When incommensurability is defended as a feature of abstract values, the rational evaluator is required to deliberate further so as to make such values more specific. This deliberative strategy is called *specification.* Its effects are analogous to the effects of commensuration, but it does not require value commensurability and is advocated in contexts marked by strong incommensurability (Richardson 1995). Second, the method of *practical induction* suitably exploits the concrete practical experience of values. This method requires that the value dimension be considered across time and allows for increasing coherence among values over time (Millgram 1997, 151–184; Millgram 2002). The evaluator learns to assess his options over time. Deliberation makes options commensurable (Millgram 1997, 157–158). Coherence is thus an achievement of deliberation, rather than a formal property of values abstractly characterized. Attention to the historical dimension of value motivates a third approach to value, which attempts to order values by situating them in the context of one's entire life. While this strategy of *life-contextualization* does not lead to any principled view of reasoning, it grounds rational action on a broad and thick conception of agential integrity. If values fit together in the context of an entire life, this is a life with integrity (Taylor 1997, 179–180).

Practical integrity is also advocated as a moral standard. For Christine M. Korsgaard decisions should be respectful of the identities that we are *willing to reflectively endorse* as practical, that is, as those identities under which we attribute ourselves values. Such practical identities may clash, on some particular occasions, and the only guide we have from practical reasoning is that we must act on the basis of reasons that everybody can share. This requirement, akin to Kant's universalization, blocks the reasons that are immoral. This is to say that moral obligations rule out contrary considerations. However, this rational guide does not prevent the possibility of severe practical conflicts among special obligations that are rooted on our identities (Korsgaard 1996, ch. 4; Scanlon 2014). Finally, one may regard rational judgment as the locus of assessment of the normative relations among values. This strategy requires that we abandon the view that commensurability and incommensurability

are hypotheses on the nature of value, and deal with them as *constitutive acts of evaluations*. Incommensurability is not a property of value that constrains reasoning, but the result of a judgment of comparative assessments, which articulates and organizes one's reasoned choice. Conversely, commensurability is the output of a successful rational deliberation, rather than its condition of possibility. Incommensurability is not an ontological feature of the value domain but a practical problem, which can be solved by engaging in deliberation. This approach to value requires a more complex view of deliberation.

As for normative determinacy, there are two related issues at stake. First, it is questionable that normative determinacy is a dominant requisite or a *desideratum* of any value theory. Second, it is questionable that commensurability and a fortiori comparability are sufficient to grant completeness in value ranking and normative determinacy of reasons for action. In other words, it is questionable that to determine what an agent ought to do we have to admit commensurability, and it is also questionable that commensurability suffices to determine what an agent ought to do. The case revolves around the relevance of so-called symmetrical dilemmas. Suppose Abe ought to financially support the synagogue and ought to financially support the museum, but he cannot afford donating to both institutions. If the deontic operator "ought" is agglomerative, then Abe ought to support either institutions, and he cannot support both, hence, he faces a dilemma. One may argue that Abe has only a disjunctive obligation: He chooses rationally if he chooses to support one of the two institutions (Herman 1993, 159–173). It would be irrational for him not to support either, but it is rational to support either one and it does not matter which one. But how does Abe decide which institution to support? The disjunctive obligation strategy leaves Abe with no decision procedure. More precisely, this strategy does not resolve the moral dilemmas, even though it indicates that the deliberative impasse can be overcome. Suppose Abe resolves to toss a coin (MacIntyre 1990). There are considerations of fairness that may guide this decision, but it is hardly the case that Abe's decision can be called fair. Perhaps one can say that it has a fair effect, but Abe's decision does not rest on any ground, nor is it chosen out of concern for fairness; hence, he cannot be judged as fair. In the case that the symmetrical choice concerns moral options, e.g., Abe has to choose between donating one instead of another, tossing a coin as a way to solve the conflict seems especially problematic, even when all deliberative routes have been explored. Some philosophers object that tossing a coin in such cases is an irresponsible act of self-indulgence (Rosalind 1999; Railton 1996, 153; Blackburn 1996, 129, 131). Others argue that the decision by randomization is arbitrary (Bagnoli 2006, 2013). These considerations support the view that symmetrical dilemmas are not trivial because of their symmetrical features. In fact, at least some symmetrical dilemmas are morally relevant hard choices. This shows that the monistic claim about commensurability does not warrant the sort of normative determinacy that many assume.

These cases are generally discounted as spurious or irrelevant on the assumption that, when there is no failure of commensurability, choice between symmetrical requirements is indifferent and can be determined by randomization. The appeal to randomization allows the agent to overcome a deliberative impasse, but it does not

really resolve the moral dilemma. This is because randomization fails to provide the agent with a genuine decisive reason for action since reasoning does not fully determine nor explain our actions. Acting in such context is not irrational, but it does not count as a principled decision. This sort of arbitrariness may not be immoral because it may not result in unfairness or bias. However, arbitrary decisions of this kind do not fully express agency and authorship of action. Lack of authorship is a failure of agential authority over one's own action, but it is not a sign of irrationality or immorality, nor does it show a failure of value commensurability.

8 Persons and Values

Persons stand in a special relation to values because they are both bearers and sources of value. This claim carries important normative and deontic implications. On a widespread view, persons have value insofar as they are persons, and they are sources of value insofar as they are persons. Correspondingly, there are evaluative practices and attitudes directed to persons as valuable, and evaluative activities by which persons assign values to other objects. Furthermore, insofar as persons are values, they are also sources of valid claims on others (Rawls 1980a, b, 452). In its turn, this implicates that there are constraints on how to relate to persons, how we should treat them, how we should express our feelings toward them, and so on. Persons carry a special relation to values insofar as they are both loci of value or value bearers, and also sources of values. How to explain this complex sort of relation is a matter of dispute. One focus of this dispute concerns the features that make persons distinctive sources of values. The other one concerns the sense in which persons are sources of values, if they produce, create, or recognize. This is an important principle, but it is not uncontested.

The main normative implication of the claim that persons are independent sources of valid claims, reasons, and values is that they are separate individuals. The separateness of persons is the grounding reason for prohibiting trade-offs. More generally, interpersonal comparisons violate the separateness of persons (Nozick 1974, 33). However, disproportion in number is also, and very generally, taken to be a ground for resolving conflicts. Both claims are deeply rooted in our commonsensical approach to value. The crucial case is a conflict between action that affects one person and action that affects many. In this sort of deliberation, it seems that numbers matter. There is a large agreement between normative theory and ordinary moral judgment that it is preferable to save the many, for instance. If numbers are relevant, though, it is because there are quantitative comparisons across persons. Arguably, there are separate reasons to save each person, since the moral value of each person is the same, and these reasons somehow add up and result in the obligation to save the many. If persons have equal significance, then the presence of each additional person should make a difference (Kamm 1989, 240–241).

Why are persons unique bearers of value? Arguably, this is because of some (metaphysical or natural) features. Kant's view is that persons have a distinct value

or dignity insofar as they have autonomy, which is a metaphysical property of the will and entitles them to exact respect from others (Kant 1997, 4.412; Kant 1996, 6.211–6.213, 227). A contemporary (and partial) rendering of this view is that persons are values because they are self-reflective, hence capable of rational and critical assessment (Nozick 1981; Frankfurt 1988; Korsgaard 1996). In this reading, reflexivity does not necessarily set humans apart from other animals (and other possible rational beings). For utilitarians, humans are morally significant insofar as they are sentient beings. To be a person does not add anything in terms of value and it is not a distinct category of value, and the personal identity of persons is irrelevant to the metaphysics of value (Parfit 1984). Persons have no special and distinct value insofar as they are persons, but only and to the extent that they are receptacles of utility. By contrast, for deontologist or a constructivist, the concept "person" indicates a normative status, which is warranted through practices of recognition, and it carries normative and deontic implications (e.g., moral claims and responsibilities). Among such implications there is respect for boundaries across persons. On this reading, then, the claim that persons have equal standing does not mean that they are commensurable items of equal value. For instance, the Kantian view is that equality of status entails that persons ought *not* to be treated as mere equivalents. Since persons have no equivalents, any exchange in value placed upon persons is morally prohibited (Kant 1997, 4.432).

9 Values and Emotions

Values stand in a complex relation with the most receptive aspect of practical rationality, which includes emotions, attitudes, and desires. These relations are complicated by the fact that there is no philosophical agreement about the concepts involved. Insofar as emotions are broadly understood as containing some conative states, along with desires, they have been often used in order to clarify the nature of evaluation and its motivational impact. For instance, emotivism holds that an evaluative judgment is not an assertion about a state of affairs or a property of an object, but expressive of emotions (Ayer 1936). As it appears, this schema of analysis assumes that the concept of emotion is clear enough to be able to explain the complex act of evaluation. However, there is a profound disagreement about what an emotion is, and whether it necessarily includes a conative state. The variety of emotions and the growing literature on the complex relation they bear with perception and reasoning strongly suggest that this analysis is doomed to failure.

However, there is a superficial agreement that at least some so-called moral and deontic emotions and values are interestingly connected. For sentimentalist theories, emotions and feelings are the source of moral judgment and also, with some corrective that pushes toward the general point of view, the cement of social life. This view is also supported by some evolutionary theorists, insisting that natural selection favors cooperative values and emotions of mutual support, recognition, care, and love may have a decisive, even though instrumental, role. According to the perceptualist

theory of value, emotions are like perceptual judgments, which allow us to detect value in the world. This view was first proposed by Scheler (1954), who argued that emotions are perceptions of "tertiary qualities," which depend upon facts about social relations, pleasure and pain, and natural psychological facts. A similar view is currently articulated by Mulligan and Tappolet (2000). Emotions such as love, compassion, care, and forgiveness shape widely shared notions of moral value. Some authors analyze this role in terms of "response dependence" and argue that emotions are responses that depend on the values and norms that lie at the core of the moral life (McDowell 1985; D'Arms and Jacobson 2000).

The relation between morality and the emotions is problematic (Bagnoli 2011). On the one hand, emotions stand in the way of moral conduct, insofar as they provide independent motivations that undermine or compete with moral motives. On the other hand, emotions seem to be necessary to have moral understanding. According to the ethics of virtue, emotions such as love and compassion are natural dispositions, which when properly habituated and educated develop into virtues of character (Baier 1985; Doris 2002; Slote 2010; Smith et al. 1989). For others, deontic moral emotions such as guilt and remorse attend the violation of duties, explain normative behavior, and signal the capacity to be bound by norms (Hare 1981, ch. 2). Emotions such as shame and pride, instead, seem to be crucial modes of valuing the self and expose individuals to the gaze of others, hence showing how the value of identity is profoundly influenced by social criteria of membership. Furthermore, the presence of emotions has been used to recover conflicts of values, and tensions between values of membership and individualistic values. For instance, genuine moral emotions and affections drive Huckleberry Finn against his judgment to abide by the law and turn Jim, the escaped slave, into the authorities (Bennett 1974; McIntyre 1990). Emotions such as love and compassion here convey attention to values that are sanctioned by socially enforced moral standards, revealing a more authentic moral understanding and attachment. On the realist theory about value, Huckleberry apprehends the moral values of human fellowship and freedom via emotional acquaintance. According to Mulligan (1998), emotions justify axiological judgments and beliefs, even though they are not direct perceptions of value. This approach seems well placed to explain how emotions further value and help moral life. However, there are significant ontological and epistemological objections against this view, which is criticized for overemphasizing the analogy with perceptual judgment. By contrast, appraisal theories of emotions hold that emotions contain an evaluative thought, but not necessarily a belief or a cognitive judgment about the case. A key point of relating values to emotions is to develop a theory of virtue, which takes emotions as natural dispositions that can be shaped by education and habituation into competences.

Recent debates have focused on the role of emotions in practical and epistemic reasoning. In traditional views, emotions are often regarded as disturbances and interferences. By contrast, empirical studies show that emotions positively contribute to reasoning and at various levels. Their basic function is to call attention to details of the situation that matter; hence, they work as criteria of salience, which help generate reasons for action and reasons for belief. Many have also started to notice and study

the role of values in theoretical reasoning, and to identify epistemic values (Pritchard 2007; Haddock et al. 2009; Williams 2002; Zagzebski 2004).

10 Valuing

Valuing is a complex activity, which concerns large varieties of objects, including properties, events, states of affairs, activities, practices, attitudes, and persons. In short, it seems that anything can be the object of assessment. Does this indicate that anything can be regarded as valuable and that there are no boundaries to what can be treated as value? As anticipated in Sect. 4, non-cognitivists and cognitivists differ in answering this question. However, they might agree that the activity of valuing admits to some constraints, even though they might disagree about what they are and how stringent. Non-cognitivists such as Hare (1963, 1981) and Stevenson (1979) hold that the constraints are so meager that any factual contents can be combined with a positive assessment, so as to identify value. If valuing is mainly an emotional attribution, even logical consistency may not apply. In fact, the varieties of pluralism, then, would be akin to ambivalences and other peculiarities of psychological lives. Arguably, the regularities and patterns that we register in recording the kinds of values cherished in the course of human history might be best explained by (evolutionary) psychology, rather than by logic or ontology (Gibbard 1990; Nozick 2001).

By contrast, cognitivists such as G. E. M. Anscombe argue that one cannot rationally value a saucer of mud (1957, 70) without indicating any rationale for supporting this preference. The rationale would be a characterization of the object in terms of its desirability; absent such characterization, the agent's valuing rests on no grounds and it is criticizable as irrational. Arguing toward a similar conclusion, Derek Parfit presents the case of somebody who cares equally about what happens every day of the week but lacks any concern for future Tuesday, conforming to a principle named "future Tuesday indifference" (Parfit 1984, 254). Singling out present Tuesday as the focus of one's valuing is irrational because it is arbitrary, lacking any justifying reasons.

The philosophical analysis of this complex activity highlights both rational and emotional components, but it is an outstanding question whether such components can be separate and, more importantly, whether we gain a better understanding of the practices of valuing by decomposing values in factual and non-factual components (Murdoch 2013). Borrowing from Murdoch, Hilary Putnam (2002, 28–45) has argued against the fact/value dichotomy on the ground that it is based on a poor understanding of evaluative language. This is a misunderstanding rooted in the empiricist tradition, which adopts a very narrow view of facts, and a very simplistic account of moral psychology. On the opposite view, facts and values are inevitably entangled. While non-cognitivism views description as devoid of values, Murdoch further suggests that one chief mode of assessing the world and deliberating about what to do is redescribing it. Moral agents become objective by constant efforts of attention, by which they attempt to redescribe reality as accurately as they can.

Murdoch's account vindicates a crucial aspect of valuing as a practice and activity importantly historical. The temporality of moral agency and of valuing is something that both realist and anti-realist have found hard to appreciate and explain. Valuing is not an occasional activity of human beings, and it does not appear to be something that we engage and disengage from at will. In fact, the very activity of valuing, both in its individual and social dimensions, seems to be profoundly related to the fact that we are temporal beings, rooted in the past, hooked in the present. Our valuing attitudes and preferences are sensitive to temporal constraints. Philosophers have identified several cases of temporal bias, in which agents discount the value of their options according to how they are situated in time. Our reasons for valuing seem to be driven by concerns that are sensitive to time. The practice of valuing thus intersects another philosophical debate about prudence and rationality over time. Philosophers disagree about whether there are or there should be normative criteria for assessing the rationality of our evaluative activities across time (Nagel 1979; Parfit 1984; Bratman 2009; Korsgaard 2008; Hedden 2015 ; Sidgwick 1907).

Furthermore, and more radically, philosophers have identified the temporality of agency as the main rationale for entering valuing practices and activities. Our subjective experience of life as temporally bounded connects crucially with what makes life worthwhile. Bernard Williams discusses the Makropulos' case—the case of an immortal but boring life, to show how the meaning of life is constrained by finitude. This is partly because the objects and desires that make a life worth living are finite and exhaustible (Williams 1973). Along these lines, Samuel Scheffler argues that what we care and value most—e.g., love and labor, intimacy and achievement, and solidarity—"have the status of *values* for us because of their role in our finite and bounded lives" (Scheffler 2013, 100). Current debates show that the issue of temporality and value is still to be placed on a clearly intelligible framework.

A promising approach is informed by the conviction that valuing is a rational activity, hardly reducible to the expression of preferences more or less intense. It is a complex activity not only because it involves a complex network of emotional and cognitive capacities, but also because it admits of various intertwined modes. We value in different ways, through varieties of evaluative judgments (e.g., along dimensions such as desirability, or reasonableness), attitudes (e.g., love, admiration, and respect), practices (ranking, mutual respect and recognition), and institutions (e.g., the market). The plurality of the modalities of valuing should be investigated not only by differentiating categories of values (e.g., intrinsic, extrinsic, instrumental, categorical), but also by understanding different (institutional and individual) modes in which we attribute and confer value in our life. A pluralist approach does not only allow us to recognize plurality of values, but it also encourages us to construct new arguments for comparing values and assessing the ethical limitations of institutions such as the market, beyond the traditional methods of welfare economics and traditional theories of justice, such as the cost–benefit analysis. Such a pluralist approach appears promising especially in consideration of the challenges faced by governmental institutions, which are required to take action in the presence of divisive conflicts of values and under uncertainty. For instance, in the debate of global warming, governments are required to take action under uncertainty, while appreciating

values as diverse as safety, financial interest, and moral obligations to future generations. While some emphasize the complexity of weighing lives through times (Broome 2004; Hedden 2015), others argue that weighing is not the appropriate way of approaching problems of rational choice, precisely because the contexts in which we choose are profoundly marked by value pluralism (Anderson 1993). This is one dramatic example of the account that the nature of value has a direct, practical impact in our lives, and affects not only the quality of the present life of our co-habitants, but also the future of life on the planet.

References

Adams, R.M. 2006. *A theory of virtue: Excellence in being for the good*. Oxford: Oxford University Press.

Anderson, E. 1993. *Value in ethics and economics*. Cambridge, Mass.: Harvard University Press.

Anderson, E. 1997. Practical reason and incommensurable goods. In *Incommensurability, incomparability and practical reason*, ed. R. Chang, 98–101. Cambridge, Mass.: Harvard University Press.

Anscombe, G.E.M. 1957. *Intention*. Oxford: Blackwell.

Anscombe, G.E.M. 1958. Modern moral philosophy. *Philosophy* 33: 1–19.

Aristotle. 1984. *The complete works of Aristotle*. Princeton, N.J.: Princeton University Press.

Audi, R. 2001. *The architecture of reason: The structure and substance of rationality*. Oxford: Oxford University Press.

Audi, R. 2005. *The good in the right: A theory of intuition and intrinsic value*. Princeton, N.J.: Princeton University Press.

Ayer, A.J. 1936. *Language, truth and logic*. London: V. Gollancz Ltd.

Bagnoli, C. 2000. Value in the guise of regret. *Philosophical Explorations* 3: 165–187.

Bagnoli, C. 2006. Breaking ties: The significance of choice in symmetrical moral dilemmas. *Dialectica* 60: 1–14.

Bagnoli, C. (ed.). 2011. *Morality and the emotions*. Oxford: Oxford University Press.

Bagnoli, C. (ed.). 2013. *Constructivism in ethics*. Cambridge: Cambridge University Press.

Baier, A. 1985. *Postures of the mind: Essays on mind and morals*. Minneapolis, Minn.: University of Minnesota Press.

Bennett, J. 1974. The Conscience of huckleberry finn. *Philosophy* 49 (188): 123–134.

Blackburn, S. 1984. *Spreading the world*. Oxford: Clarendon Press.

Blackburn, S. 1985. Errors and the phenomenology of value. In *Morality and objectivity*, ed. Ted Honderich. London: Routledge.

Blackburn, S. 1993. *Essays in quasi-realism*. New York, N.Y.: Oxford University Press.

Blackburn, S. 1996. Dilemmas: Dithering, plumping, and grief. In *Moral dilemmas and moral theory*, ed. H. Mason, 127–139. New York: Oxford University Press.

Blackburn, S. 1998. *Ruling passions*. Oxford: Oxford University Press.

Boyd, R. 1988. How to be a Moral Realist. In *Essays on moral realism*, ed. G. Sayre-McCord, pp 181–228. Cornell University Press.

Brandt, R.B. 1996. *Facts, values, and morality*. Cambridge University Press.

Brink, D.O. 1984. Moral realism and the sceptical arguments from disagreement and queerness. *Australasian Journal of Philosophy*, 62(2): 111–125.

Broome, J. 1991. *Weighing goods: Equality, uncertainty and time*. Oxford: Blackwell.

Broome, J. 1997. Is incommensurability vagueness? In *Incommensurability, incomparability and practical reason*, ed. R. Chang, 67–89. Cambridge, Mass.: Harvard University Press.

Broome, J. 2000. Incommensurable values. In *Well-being and morality: Essays in honour of James Griffin*, ed. R. Crisp, and B. Hooker, 21–38. Oxford: Clarendon Press.
Broome, J. 2004. *Weighing lives*. Oxford: Oxford University Press.
Broome, J. 2013. *Rationality through reasoning*. Oxford: Wiley-Blackwell.
Chang, R. 1997. Introduction. In *Incommensurability, incomparability, and practical reasoning*, ed. R. Chang. Cambridge, Mass.: Harvard University Press.
Chang, R. 2002. *Making comparisons count*. London-New York: Routledge.
D'Arms, J., and D. Jacobson. 2000. Sentiment and value. *Ethics* 110: 722–748.
Dancy, J. 2000. *Practical reality*. Oxford: Oxford University Press.
Dancy, J. 2004. *Ethics without principles*. Oxford: Oxford University Press.
Darwall, S., A. Gibbard, and P Railton (eds.). 1996. *Moral discourse and practice*. Oxford University Press USA.
Deigh, J. 2008. *Emotion, values, and the law*. New York, N.Y.: Oxford University Press.
Doris, J. 2002. *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.
Dworkin, R. 1977. Taking rights seriously. *Philosophical Quarterly* 27 (109): 379–380.
Finnis, J. 1980. *Natural law and natural rights*. Oxford: Oxford University Press.
Firth, R. 1951. Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research* 12 (3): 317–345.
Foot, P. 2000. *Natural goodness*. Oxford: Clarendon Press.
Frankena, W.K. 1939. The naturalistic fallacy. *Mind* 48: 464–477.
Geach, P. 1956. Good and evil. *Analysis* 17: 33–42.
Geach, P. 1960. Ascriptvism. *Philosophical Review* 69: 221–225.
Gibbard, A. 1990. *Wise choices, apt feelings: A theory of normative judgment*. Harvard University Press.
Gibbard, A., and A. Macintyre. 1995. The viability of moral theory. *Philosophy and Phenomenological Research* 55: 343–356.
Griffin, J.1986. *Well-being: Its meaning, measurement and moral importance*. Clarendon Press.
Haddock, A., A. Millar, and D. Pritchard (eds.). 2009. *Epistemic value*. Oxford: Oxford University Press.
Hare, R.M. 1952. *The language of morals*. Oxford: Clarendon Press.
Hare, R.M. 1963. *Freedom and reason*. Oxford: Clarendon Press.
Hare, R.M. 1981. *Moral thinking: Its levels, method, and point*. Oxford: Oxford University Press.
Harman, G. 1977. *The nature of morality: An introduction to ethics*. Oxford: Oxford University Press.
Harman, G. 2000. *Explaining value and other essays in moral philosophy*. Oxford University Press.
Harman, G., and J.J. Thomson. 1996. Moral relativism and moral objectivity. *Philosophy* 71: 622–624.
Hedden, B. 2015. *Reasons without persons: Rationality, identity, and time*. Oxford University Press UK.
Herman, B. 1993. *The practice of moral judgment*. Harvard University Press.
Hume, D. 1739. *A treatise of human nature*. Oxford University Press.
Hurka, T. 1993. *Perfectionism*. New York: Oxford University Press.
Rosalind, H. (1999). *On virtue ethics*. Oxford University Press.
Hurka, T. 2000. *Virtue, vice and value*. New York, N.Y.: Oxford University Press.
Joyce, R. 2001. *The myth of morality*. Cambridge University Press.
Kamm, F. 1989. Harming some to save others. *Philosophical Studies* 57: 227–260.
Kant, I. 1996. *Metaphysics of morals*, trans. Mary Gregor, 1st ed, 1797. Cambridge: Cambridge University Press.
Kant, I. 1997. *Groundwork for the orals*, trans. Mary Gregor, 1st ed, 1785. Cambridge: Cambridge University Press.
Kolodny, N. 2003. Love as valuing a relationship. *Philosophical Review* 112: 135–189.
Kolodny, N. 2005. Why be rational? *Mind* 114: 509–563.

Korsgaard, C. 1983. Two distinctions in goodness. *Philosophical Review* 92: 169–195.
Korsgaard, C. 1996. *The sources of normativity*, ed. O. O'Neill. Cambridge: Cambridge University.
Korsgaard, C. 2008. *The constitution of agency: Essays on practical reason and moral psychology*. Oxford: Oxford University Press.
Lukes, S. 1997. Comparing the incomparable: Trade offs and sacrifices. In *Incommensurability, incomparability and practical reason*, ed. R. Chang, 185–186. Cambridge, Mass.: Harvard University Press.
Lyons, D. 1976. Ethical relativism and the problem of incoherence. *Ethics* 86: 107–121.
MacIntyre, A. 1988. *Whose justice? Which rationality?* Notre Dame, Minn.: University of Notre Dame Press.
MacIntyre, A. 1990. Moral dilemmas. *Philosophy and Phenomenological Research* 50: 367–382.
MacIntyre, A. 2007. *After virtue: A study in moral theory*. Notre Dame, Minn.: University of Notre Dame Press.
MacIntyre, A. 2008. Value and context: The nature of moral and political knowledge. *Journal of Moral Philosophy* 5: 151–154.
Mackie, J.L. 1977. *Ethics: Inventing right and wrong*. New York, N.Y.: Penguin.
McDowell J. 1985. Values and secondary qualities. In *Morality and Objectivity*, ed. Ted Honderich. London: Routledge. pp. 110–129.
Mill, J.S. 1988. Utilitarianism. In *Collected works of John Stuart Mill*, 1st ed, 1861, vol. 29, ed. J.M. Robson, 371–577. Toronto: University of Toronto Press.
Millgram E. 1997. *Practical induction*. Harvard University Press.
Millgram E. 2002. Commensurability in perspective. *Topoi* 21 (1–2): 217–226.
Moore, G.E. 2004. *Principia Ethica*, 1st ed, 1903. Mineola, NY: Dover Publications.
Moore, G.E. 2010. *Philosophical studies*, 1st ed, 1921. London: Routledge.
Mulligan, K. 1998. From Appropriate emotions to values. *Monist* 81: 161–188.
Murdoch, I. 2013. *The Sovereignty of good*, 1st ed, 1971. London: Routledge.
Nagel, T. 1979. *The possibility of altruism*. Princeton: Princeton University Press.
Nagel, T. 1986. *The view from nowhere*. Oxford: Oxford University Press.
Nichols, S. 2004. *Sentimental rules: On the natural foundations of moral judgment*. Oxford: Oxford University Press.
Nozick, R. 1974. *Anarchy, State and Utopia*. Oxford: Blackwell.
Nozick, R. 1981. *Philosophical explanations*. Cambridge, Mass.: Harvard University Press.
Nozick R. 2001. *Invariances: The structure of the objective world*. Belknap Press of Harvard University Press
Nussbaum M.C. 1990. *Love's knowledge*. Oxford University Press.
O'Neill, O. 1989. *Constructing authorities: Reason, politics and interpretation in Kant's philosophy*. Cambridge: Cambridge University Press.
Parfit, D. 1984. *Reasons and persons*. Oxford: Oxford University Press.
Parfit, D. 2006. Normativity. *Oxford Studies in Metaethics* 1: 325–380.
Parfit, D. 2011. *On what matters*. Oxford: Oxford University Press.
Plato. 1991. *Euthyphro*, ed. Chris Emlyn-Jones. Bristol: Bristol: Bristol Classical Press.
Prinz, J.J. 2007. *The emotional construction of morals*. Oxford: Oxford University Press.
Pritchard, D. 2007. Recent work on epistemic value. *American Philosophical Quarterly* 44: 85–110.
Putnam, H. 2002. *The collapse of the fact/value dichotomy and other essays*. Cambridge, Mass.: Harvard University Press.
Railton 1996, The Diversity of moral dilemma. In *Moral Dilemmas and Moral Theory*, ed. Mason, H. E. (1996). Oxford University Press.
Rawls, J. 1980a. Construction and Objectivity. *Journal of Philosophy* 77 (9): 554–572.
Rawls, J. 1980b. Kantian constructivism in moral theory. *Journal of Philosophy* 77 (9): 515–572.
Rawls, J. 1993. *Political liberalism*. Columbia University Press.
Rawls, J. 1994. Political Liberalism. Philosophical Quarterly 44 (177):542-545.
Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.

Richardson H.S. 1995. Beyond good and right: Toward a constructive ethical pragmatism. *Philosophy and Public Affairs* 24 (2): 108–141.
Richardson, H.S. 1999. Institutionally divided moral responsibility. *Social Philosophy and Policy* 16 (2): 218.
Richardson, H.S. 2013. Moral reasoning. *The Stanford Encyclopedia of Philosophy*.
Ross, W. D. 1930. *The right and the good*. Clarendon Press.
Scanlon, T.M. 1991. The Moral basis of interpersonal comparisons. In *Interpersonal comparisons of well-being*, ed. Jon Elster and John E. Roemer, 17–44. Cambridge University Press.
Scanlon, T.M. 1995. Moral theory: Understanding and disagreement. *Philosophy and Phenomenological Research* 55: 343–356.
Scanlon, T.M. 1998. *What we owe to each other*. Cambridge, Mass.: Belknap Press of Harvard University Press.
Scanlon, T.M. 2014. *Being realistic about reasons*. Oxford: Oxford University Press.
Scheler, M. 1954. *The nature of sympathy*. Hamden, Conn.: Archon.
Scheffler, S. 2013. *Death and the afterlife*. Oup Usa.
Schroeder, M.A. 2009. *Slaves of the passions*. Oxford: Oxford University Press.
Schroeder, M.A. 2010. *Noncognitivism in ethics*. London: Routledge.
Schroeder, M.A. 2011. Moral sentimentalism. *Philosophical Review* 120: 452–455.
Sen, A., and B. Williams. 1982. Introduction: Utilitarianism and beyond. In *Utilitarianism and beyond*, ed. A. Sen, and B. Williams. Cambridge: Cambridge University Press.
Shafer-Landau, R. 1994. Ethical disagreement, ethical objectivism and moral indeterminacy. *Philosophy and Phenomenological Research* 54: 331–344.
Sidgwick, H. 1907. *The methods of ethics*. Indianapolis, Ind.: Hackett.
Skorupski, J. 1993. The definition of morality. *Royal Institute of Philosophy Supplement* 35: 121–144.
Skorupski, J. 2010. *The domain of reasons*. Oxford University Press.
Slote, M. 1989. *Beyond optimizing*. Cambridge, Mass: Harvard University Press.
Slote, M. 1992. *From morality to virtue*. Oxford: Oxford University Press.
Slote, M. 2010. *Moral sentimentalism*. Oxford: Oxford University Press.
Smith, M., D. Lewis, and M. Johnston. 1989. Dispositional theories of value. *Proceedings of the Aristotelian Society* 63: 89–174.
Smith, M. 1994. *The moral problem*. Oxford: Blackwell.
Stevenson, C.L. 1937. The emotive meaning of ethical terms. *Mind* 46: 14–31.
Stevenson, C.L. 1963. *Facts and values*. New Haven, Conn.: Yale University Press.
Stevenson, C.L. 1979. *Ethics and language*. Norwalk, Conn.: Ams Press.
Stevenson, C.L. 2009. The nature of ethical disagreement. In *Exploring philosophy: An introductory anthology*, ed. S.M. Cahn. Oxford: Oxford University Press.
Stocker, M. 1989. *Plural and conflicting values*. Oxford University Press.
Stocker, M. 1998. Emotions. How emotions reveal value and help cure the schizophrenia of modern ethical theories. In *How should one live?: essays on the virtues*, ed. Roger Crisp. Clarendon Press.
Sturgeon, N. 1998. Moral explanations. In *Ethical theory 1: The question of objectivity*, ed. James Rachels. Oxford University Press.
Sturgeon, N.L. 1986. Harman on moral explanations of natural facts. *Southern Journal of Philosophy* 24 (S1): 69–78.
Tappolet, C. 2000. *Emotions et Valeurs*. Paris: Presses Universitaires de France.
Taylor, C. 1997. Leading a Life. In *Incommensurability, incomparability, and practical reasoning*, ed. R., Chang. Cambridge, Mass.: Harvard University Press. chapter 9.
Taylor, C. 1976. Responsibility for self. In *The Identities of persons*, ed. Amelie oksenberg rorty, pp. 281–299. University of California Press.
Taylor, C. 1989. *Sources of the self: The making of the modern identity*. Harvard University Press.
Thompson, M. 2008. *Life and action: Elementary structures of practice and practical thought*. Cambridge, Mass.: Harvard University Press.
Toulmin, S.E. 1950. *An examination of the place of reason in ethics*. Cambridge University Press.

Velleman, J.D. 2009. *How we get along*. Cambridge: Cambridge University Press.
Velleman, J.D. 2013. *Foundations for moral relativism*. Cambridge: OpenBook Publishers.
Wiggins, D. 1987. *A sensible subjectivism?*. Blackwell.
Williams, B. 1973. The Makropulos case: Reflections on the tedium of immortality. In Id., *Problems of the Self.* Cambridge: Cambridge University Press.
Williams, B. 1981. *Moral luck*. Cambridge: Cambridge University Press.
Williams, B. 1985. *Ethics and the limits of philosophy*. Cambridge, MA: Harvard University Press.
Williams, B. 2002. *Truth and truthfulness: An essay in genealogy*. Princeton, N.J.: Princeton University Press.
Wong, D. 1986. On moral realism without foundations. *Southern Journal of Philosophy* 24: 95–113.
Wong, D. 1989. Three kinds of incommensurability. In *Relativism: Interpretation and confrontation*, ed. M. Krausz, 140–158. Notre Dame, Ind.: Notre Dame University Press.
Wong, D. 1992. Coping with moral conflict and ambiguity. *Ethics* 102: 763–784.
Wong, D. 2006. *Natural moralities: A defense of pluralistic relativism*. Oxford: Oxford University Press.
Wong, D. 2008. Constructing normative objectivity in ethics. *Social Philosophy and Policy* 25: 237–266.
Zagzebski, L. 2004. Epistemic value and the primacy of what we care about. *Philosophical Papers* 33: 353–377.

The Goals of Norms

Cristiano Castelfranchi

Premise: The Foundational and Intrinsic Relation between N and G

The relationship between norms (Ns) and goals is not simply important but foundational, since *Ns are artifacts for social coordination through agents' goal manipulation*.

This relation is also multifaced and multifunctional; these faces and functions—in order to understand what Ns are and how they work—should be explicitly analyzed and explained.

Obviously, our object is "norms" in the "normative" (prescriptive) meaning/sense, not in the "normality" (descriptive or statistic or standard sense). However, there is an important and bidirectional goal relation between N1 (in the normative sense) and N2 (in the normality sense):

(a) N2 creates and becomes a goal for the actors and even an N1 (a prescription, something "due"), in order to conform, to be like others. This conformity is either a need of the individual or a need (and request/pressure) of the group, or both.
(b) N1 creates an N2, a normal conduct in the community, if it is respected: N conformity is "normal."

Moreover:

(b1) N1 has the goal and function to be respected and thus to create an N2, a normal behavior (at the individual, internal level this helps it to also become an automatic response, just a habit);
(b2) If N1 does not become/create an N2, it is weakened (Bicchieri 2006; Conte and Castelfranchi 1995), perceived as less credible and less binding.

C. Castelfranchi (✉)
Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche (CNR), Rome, Italy
e-mail: cristiano.castelfranchi@istc.cnr.it

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_7

Coming back to Ns as behavior regulation devices, we have:

On the one side, Ns are *only* for autonomous *goal-directed* agents, whose behavior depends on their goals/choices, preferences, that is, on their free decisions, and are thus assumed to be "responsible" for their action, since they could behave differently.

On the other side, norms "have goals" in the sense that they are built and used "for" something. They are interactional and societal *tools*.

But they have goal and are finalistic in another sense, too: They guarantee certain social emerging *functions*. This is a different teleological notion and a different kind of "goal." It is crucial to disentangle these notions and to have a theory of their structural and functional relations.

A good theory (and study) of Ns should be explicit in distinguishing and factually exploring the differences and relations between:

1. The intention (intended effects) of N; what we expect and want to produce by the issuing and monitoring a given N;
2. The actual effects of N, and in particular:

   2a. *N efficacy*, that is, the actual/realized intended effect (corresponding to its goals or unsatisfactory); such *efficacy* or failure clearly is evaluated with respect to the "goal" of N: (1).

   2b. *Functional effects*, usually unintended (desirable or even undesirable) and not understood, but such that they have feedback and select and reproduce that behavior or entity, including a given N; the outcome (in part) is responsible for the maintenance or reproduction of N.[1]

   2c. *Side effects*, neither intended nor functional but systematic, and either negative (as with addictive drugs) or positive (relative to some other goals not in the plan of N).

In other words, Ns have to do with *two kinds of teleology*: (i) mentally represented (and perhaps even intended) or *psychological goals* that regulate our conduct and (ii) nonmental goals, merely emergent and self-organizing "functions" (social or biological) impinging on our individual and collective behaviors. Let us use the term *goal* only for the internal control system, the mentally represented objective, and the term *function* only for the external selecting finality of a feature or behavior.

Given this distinction between two kinds of finality impinging on norms, the question is: What are the dialectics between them? More precisely:

(a) Is the goal of N the goal (G) represented in the mind of "subject" S, and driving her behavior? Is it at least the G in the mind of the "issuer"?
(b) Should S understand and pursue the G of the N while obeying it?

As for (a), the answer is: only partially; the goal induced in the S's mind and adopted by her is only a subgoal of N, not its aim in the mind of the authority nor its social function.

[1] Consider, for example, Marx's claim that prisons also reproduce themselves by reproducing delinquency; which of course is not the mission or the aim of prisons! But in a sense is a "bad *function*" of them.

And as for (b), the answer is: not at all; on the contrary, the ideal "obedience" and subject are without a cooperative sharing of the aim of N and in any case are not motivated by such explicit collaboration.[2]

So we have to examine what a goal is; what a goal-directed behavior is; how goals regulate human behavior; how Ns are *made to* induce goals and regulating our behavior, but also what the "functions" are and what the relation is between the goals of the "subject," the goal imposed by N in the subject's mind, the issuer's goal, and the function of N.

1 Goals–Norms: A Multiple Relation

Six are the main structural relations between Ns and Gs:

(A) Ns are communicative artifacts for "social manipulation," designed to influence autonomous, goal-directed behavior *by inducing or blocking goals* in a given set of addresses. *Ns are goal-oriented tools for the goal-directed behavior of autonomous agents.* They want to enter our preferences and decisions or bypass them by building a habit, an automatic conform response.

(B) Ns presuppose and need the *postulation of goals in the minds of subjects*; they are grounded in an "intentional stance" and the attribution of mental states and in particular of autonomous behavior internally regulated by goals and beliefs and then by "choices/decisions."[3]

(C) Ns are aimed at governing our conduct through our goal mechanism: at "regulating" our behavior by "influencing" us, that is, by *inducing* goals in us, goals to be *adopted*, so that they become internal and possibly prevail on personal/private motives.
Either by inducing us to do something or by creating a conflict and inhibiting a "wrong" behavior on our part. Norms *reduce* the subject's choices—the possible goals to be pursued—and/or *add* new goals to be pursued.
They have to "give" us new goals (or new reasons for a goal) and to block other possible goals we may have. Finally, they have to generate the *intention* regulating our action.

(D) Ns are for goal adoption (*adhesion*). Since we are autonomous agents (governed and motivated by our own (internal) motives), the N-goal must be internalized and adopted for some goal we may have. But N has the goal (and function) and (cl)aim to be adopted for specific goals and reasons. The ideal-typical adhesion to an N is for an intrinsic motivation, for a "sense of duty," recognition of

[2]When driving a car, you "have to" use a safety belt even if you disagree about this prescription and use: It is not the case that you have to use one on condition that you agree that it is better for you or for the costs placed on the community.

[3]It is useless to create "norms/laws" for animals; better to use orders and threats, or physical barriers. In humans, mental barriers can work, and frequently even physical barriers primarily have a signaling, communication, function.

authority, because it is right/correct to respect Ns. Only subideally should one respect Ns, in order to avoid external or internal sanctions. Normative education also goes in this direction.
Only after the goal adoption, there is "true" (intentional/aware) "violation" (disobedience) not just a behavioral violation.
Eventually, norm-conforming behavior can be proceduralized, automatized, routinized; however, on the one side, there is the implicit knowledge that it is a norm; on the other side, it socially counts *"as if"* were a deliberate and intentional act (responsibility ascription).

(E) Ns work thanks to the distributed, collective, expectation about others' goal of conformity, that is, a collective and mutual "prescription" (*goals about the goals* of others), and monitoring and sanctioning.

(F) Norms are "tools for" something: *They have a goal, and they are aimed at producing a given result* (coordination of actions and interests, social order, power distribution, reference rules, and trust). They have a "goal" (the goal of the "legislator," of the normative authority, possibly to be also understood and interpreted by the judge and the monitoring guy but not necessarily understood or intended by the subject (see later)) and some function. The "function" of Ns is not necessarily understood and intended, and not necessarily a good one and corresponding to the norm's official end/mission.

We will try to illustrate these faces of the Ns–goals relation.

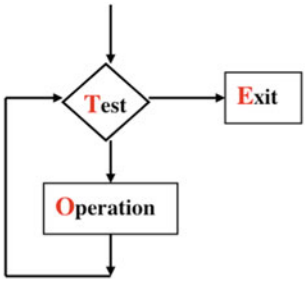
2 Teleologies of Mind: Goals, Functions, and Pseudogoals

2.1 What Are Goals?

In modern science, there are two well-defined teleological frames and notions:

- The one provided by *evolutionary approaches*, where it is standard (and correct) to speak in terms of functions, (adaptive) value, being *for* something, having a certain finality/end, providing some advantage, etc. In this context, *goal* (end, function, finality, etc.) means the "effect" (outcome) that has selected/reproduced and maintained a certain feature or behavior: initially just an accidental effect, an effect among many others, but later, thanks to the loop and positive feedback on its own "causes" (i.e., on the feature or behavior producing it) no longer a mere effect but the "function," the purpose of that feature, what makes it useful and justifies its reproduction.
- The one provided by *cybernetic control theory* and its postulated and representations cycle, in which the agent is able to adjust the world through goal-directed behavior, and to maintain a given "desired" state of the world (homeostasis), as illustrated in Fig. 1.

Fig. 1 TOTE cycle



Actually, there might be a third teleological/finalistic notion used in several sciences (from medicine to the social sciences): the notion of a "function" of X as a "role," a functional component, an "organ" of a global "system"; for example the "function" of the heart or of the kidneys in our body, or the function of families (or of education or of norms) in a society, or the function of a given office in an organization. However, this "functionalist" and "systemic" notion has never been well defined and has elicited a lot of problems and criticisms. My view is that this finalistic view is correct, but it is reducible to, and derived from, the previous two kinds of teleology. Either the "organs" are the result of an evolutionary selection—in that they contribute to the fitness and reproduction (maintenance) of that organism—or there is a "project," a "design," that is, a complex goal in someone else's mind, a goal which imposes particular subgoals on its parts, components, and tools, or both.

2.2 The Relations Between Psychological Goals and Behavior Functions

A serious problem for a (future) science of goals is that these two fundamental teleological notions/mechanisms have never been unified:

i. Neither conceptually, by looking for a common definition, a conceptual common kernel (e.g., in terms of circular causality, feedback): Do we have and is it possible to have a general, single notion of "goal" with two subkinds (functions vs. psychological goals)?
ii. Nor by solving the problem of the interaction between the two coexisting forms of finality.

This constitutes a serious obstacle and reveals a real ignorance gap in contemporary science.[4]

[4] As for issue (i), without the aforementioned conceptual unification, we cannot have a unitary theory of communication—or a theory of cooperation, of sociality, etc.—in animal and humans. What are today presented as unified theories are just a trick. In fact, these notions—which necessarily require a goal (e.g., "communication" requires not only a sign-"reader" but also a "sender": The information

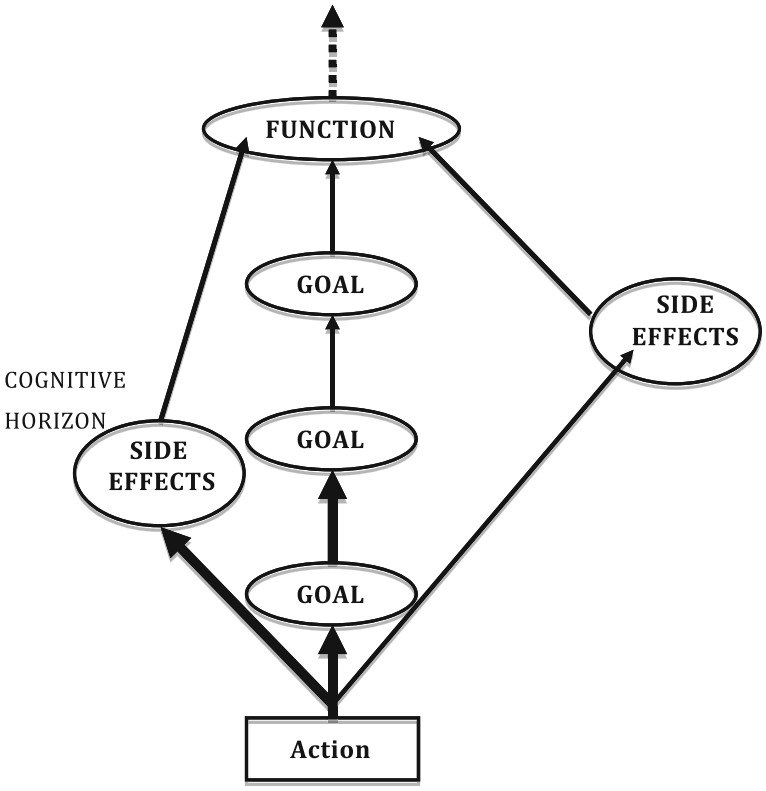


Fig. 2 Mental goals and possible functions

As for point (ii): *What is the relationship between the internally represented goals (motivations and concrete objectives) of an agent regulating its behaviors from the inside and the adaptive functions that have selected that agent and its behaviors?*

Usually, in purposive, goal-driven agents/systems, the "function" of their conduct, the adaptive result that has to be guaranteed, is *not* internally represented and pursued; it is neither understood nor foreseen (Fig. 2). Of course, not all foreseen outcomes or all side effects have a "function."

The internal motivations (and whatever solutions and instrumental goals they generate) are just subgoals of the "external" goals of the behavior, of its functions;

is deliberately "given" to the "receiver/addressee")—are defined in terms of adaptive functions when applied to simple animals (like insects), whereas in humans they are defined in intentional terms. Thus, there is no unified notion (or theory) of "communication," in that we do not know the common kernel between a "functional" device and an "intentional" device. A remarkable attempt to deal with these problems is Ruth Millikan's work.

they are just "cognitive mediators" of the (biological or social) functions that would be nonrepresentable and mentally noncomputable.[5]

2.3 Goals Versus Pseudogoals

It is also very important to disentangle true goals from *pseudogoals* (Castelfranchi 2012), that is, goals that only seem to be there and to regulate the system and its behavior. However, they are not in fact there as goal mechanisms: They are not represented in and "governing" the system. They are just functional ways in which the system has been "designed" (by evolution, by learning, by the designer); they are the system's goal-oriented way of working, its operational rules. For example, a real thermostatic system (thermostat, thermometer, room, radiator, boiler, etc.) has been designed in order to reduce naphtha consumption, heat loss, etc., as much as possible. These are (pseudo) goals of the system, which also works to guarantee them, but they are not true cybernetic goals like the thermostat's set point. They are not represented, evaluated, and "pursued" by the system action cycle.

Analogously, our minds have been shaped (by natural selection, or culture and learning) in order to have certain working principles and to guarantee certain functions, which are not explicitly represented and intended. It seems (from our behavior) that we have certain goals, but they are not real goals, only pseudogoals. This is the case, in our view, with some well-known (and badly misunderstood) finalistic notions, like utility maximization, cognitive coherence, and even pleasure. No doubt, we often choose between different possible goals so as to maximize our expected utility, giving precedence/preference to the greater expected value, that is obvious and adaptive. However, this does not mean that we have *the* goal (the single and monarchic goal) of maximizing our utility, regardless of specific contents and goods. On the contrary, we are moved and motivated by specific, qualitative terminal goals we want to achieve (esteem, sex, power, love, etc.), but the *mechanism* that has to manage them has been designed and works so that it maximizes expected utility.

In the same vein, we maintain coherence among our beliefs, and need to avoid and eliminate contradictions. That is why we can reject certain information and do not believe all the data we get (sometimes even what we directly perceive; "we do not believe our eyes," literally); the new data must be "plausible," credible, integrable, within the context of our preexisting knowledge; otherwise, we have to revise our previous beliefs on the basis of new (credible) data. This coherence maintenance is frequently completely automatic and routine. We have mechanisms for checking and

[5] For example, only very recently have we discovered why we have to eat, the real functions/effects of our food in our organisms (proteins, carbohydrates, vitamins, etc.), and very few people eat in view of such effects. We eat for hunger or pleasure or out of habit. Analogously, we do not usually engage in courtship and sex in view of reproduction: We are driven by other internal motives. We can even cut off the "adaptive" connection between our motives and their original functions, as by deciding to have sex without inseminating or without establishing/maintaining any friendly/affective relation or support.

adjusting coherence. We do not usually have any real intention about the coherence of what we believe. Thus, knowledge coherence is a pseudogoal of ours, not a real meta-goal guiding meta-actions.

2.4 Subjective Kinds of Goals

- Neither *goals* nor *motives* mean "desires." Desires are just one kind of goal. Desires are endogenous (and usually pleasant), and with Ns, we just have to cut off some possible course of action by *making some desire of the S practically impossible or nonconvenient*. It is ignored that *"duties" are not "desires";* they are *goals from a different source*, with a different origin: They come from the outside (*exogenous*),[6] and they are imported, "adopted"; they are "prescriptions" and "imperatives" from another agent.
- Not all goals have to be "actively pursued" or just "pursued"; some of them (like having a sunny day) are not within our power: Whether they can be realized is not up to us but depends on other "agents" or external forces, so we cannot really "pursue" them. Other goals are such that their realization is only partly up to us: We have something to do with it, but then the final result depends on others or on luck, an example being winning a lottery or being acquitted of a crime we did not commit. As was previously pointed out, a goal is not a goal only if/when actively pursued.
  Thus, we may have *actively* pursued goals (goals pursued through our active actions), but also merely *passive* goals, and the latter can be of two very different kinds:
  - Goals we can only wait for, hoping that they will be attained, for they do not depend on us at all: We cannot do anything (else).
  - Goals whose realization depends on us and on our "doing nothing," that is, on our abstaining from possible interference. We would have the power to block that event/result, and we decide to do nothing, in order to let it happen (inaction, "passive action"). This case also involves our "responsibility," since the result is due to our decision and (in)action.
- An important distinction in motivational theory is that between avoidance goals and approach goals. This is how Wikipedia summarizes the difference: "Not all goals are directed towards *approaching* a desirable outcome (e.g., demonstrating competence). Goals can also be directed towards *avoiding* an undesirable outcome" (for scholarly discussion, see, for instance, Elliot 2006). More than that, avoidance and approach represent two mental frames, two different psychological

[6]However, see the later discussion on the internalization of "authority" and on internal moral imperatives.

dispositions and mind-sets (see Higgins's 1997 avoidance and approach "regulatory *focus*").[7]

In avoidance goals, success is to pass from Q to not Q (to end the negative state) or to prevent passage from not Q to Q, while in approach goals, success is to pass from not P to P or from P to P (i.e., maintaining P as the desired status quo).[8]

- Not all our goals are "felt," in part because not all of them are represented and defined in a sensory-motor format.[9] The two most important kinds of felt goals are *desires* and *needs*.

  In the "felt need" for a given object O, we perceive a current, unpleasant, or disturbing bodily or affective stimulus S (for instance, we perceive dryness in our throat when we "*feel* the need for water") that we cognitively ascribe to the loss of O. Similarly, in felt desires, we just "imagine" and anticipate the pleasant sensations/emotions that we will/would have if/when attaining our cherished object.
- Intentions are those goals that *actually drive our voluntary actions or are ready/prepared to drive them*. They are not another "primitive" (like in the BDI model inspired by Bratman's theory: see, e.g., Rao and Georgeff 1995), a mental object different from goals. They are just a kind of goal: the final stage of successful goal processing, which also includes "desires" in the broad sense,[10] with very specific and relevant properties (see also Castelfranchi and Paglieri 2007; Castelfranchi et al. 2007).

In a nutshell, in our model, an *intention* is a goal that

(1) has been activated and processed;
(2) has been evaluated as not impossible, and not already realized or self-realizing (achieved by another agent), and whose achievement is therefore *up to us*: We have to act in order to achieve it[11];

---

[7] Notice how this terminology (e.g., "approach") is related to a semantics/connotation of "goal" that was criticized in Sect. 3, point A, as being too strongly inspired and constrained by behaviorism. Moreover, many motivational theories about avoidance and approach (such as Higgins's) remain essentially hedonistic.

[8] Another important difference is between gradable and all-or-nothing goals, or between achievement and maintenance goals. But these distinctions here are less relevant (Castelfranchi 2012).

[9] This means that we cannot say, for example, "I feel the intention of…"—for the simple reason that sensory-motor format of the represented anticipatory state is not specified in the notion of intention. Intention is a more "abstract" notion of goal, with an unspecified codification. Looking at a goal as an "intention," we abstract away from its possible sensory components.

[10] The creation of two distinct "primitives," basic independent notions/objects (desires vs. intentions), is in part due to the wrong choice of adopting "desires" (also in accordance with common sense) as the basic motivational category and source. We have already criticized this reductive move (Sect. 4) and introduced a more general and basic (and not fully common sense) teleonomic notion. This notion also favors a better unification of kinds of goals and a better theory of their structural and dynamic relationships.

[11] An intention is always an intention to "do something" (including inactions). We cannot really have intentions about the actions of other autonomous agents. When we say something like "I have the intention that John go to Naples," what we actually mean is "I have the intention *to bring it about that* John goes to Naples."

(3) has been chosen against other possible active and conflicting goals, and we have "decided" to pursue it;
(4) is consistent with other intentions of ours; a simple goal can be contradictory, inconsistent with other goals, but once chosen, it becomes an intention and has to be coherent with our other intentions (Castelfranchi and Paglieri 2007; Castelfranchi et al. 2007)[12];
(5) implies the agent's belief that she knows (or will/can know) how to achieve it, that she is able to perform the needed actions, and that there are or will be the needed conditions for the intention's realization; at least, the agent believes that she will be able and in a condition to "try";
(6) is being "chosen" implies a "commitment" with ourselves, a mortgage on our future decisions; intentions have priority over new possible competing goals and are more persistent than the latter (Bratman 1987);
(7) is "planned"; we allocate/reserve some resources (means, time, etc.) to it, and we have formulated or decided to formulate a plan consisting of the actions to be performed in order to achieve it. An intention is essentially a two-layer structure:

   (a) the "intention *that*," the *aim,* that is, the original processed goal (e.g., to be in Naples tomorrow) and
   (b) the "intention *to do*," the subgoals, the planned executive actions (to take the train, buy the tickets, go to the station, etc.). There is no "intention" without (more or less) specified actions to be performed, and there is no intention without a motivating outcome of such action(s);

(8) thus, an intention is the final product of a successful goal processing that leads to a goal-driven behavior.

After a decision to act, an intention is already there even if the concrete actions are not fully specified or are not yet being executed, because some condition for its execution is not currently present. Intentions can be found in two final and prefinal stages:

(a) *intention "in action,"* that is, guiding the executive "intentional" action;
(b) *intention "in agenda"* ("future-directed," those more central to the theories of Bratman, Searle, and others), that is, already planned and waiting for some lacking condition for their execution: time, money, skills, etc. For example, I may have the intention to go to Capri next Easter (the implementation of my "desire" of spending Easter in Capri), but now it is February 17, and I am not going to Capri or doing anything to that end; I have just decided to do so at the right moment; it is already in my "agenda" ("things that I have to do") and binds my resources and future decisions.[13]

[12]Decision-making serves precisely the function of selecting those goals that are feasible and coherent with one another, and allocating resources and planning one's actual behavior.

[13]I would also say that an "intention" is "conscious": We are aware of our intentions, and we "deliberate" about them; however, the problem of unconscious goal-driven behavior is open and quite complex (see Bargh et al. 2001).

Ns are aimed at producing (through a choice) *intentions* in the subject; on the basis of given beliefs: "Is it really an N? Is it valid and respected? Is it my case?… Are there other conflicting, more important Ns?… Am I able and in the condition?…"

This is the cognitive *N-processing* in minds (Castelfranchi 2013; Conte et al. 2010).

3 Features of the Goal of an N

3.1 Impersonal

The goal of "ordering"/"prohibiting," of issuing an imperative, cannot be personal/private; you have to play your role and to implement, or actuate a goal of the role; so your will cannot to be your will but an "institutional" normative will.

It is the goal "of the authority," that is "who" we consider "*entitled*" (meta-norm) to issue an N: God; the group or community, the impersonal anonymous "we" and "one," "nobody," etc.; the boss or leader; the institutionalized authority (the king, Parliament), and an intrinsic part of acknowledging and treating that impinging will as an N (and thus as a deontic obligation) is to recognize the goal as a nonpersonal and rightful one.

So an N and a simple prescription, imperative, or impositive request differ not by the "object" ("Do not smoke," "Close the door!") by for the required motivation, and so, in a sense, by their "content," since it is part of the content of the "linguistic" (or communicative) normative act (issuing or instancing N) to prescribe specific motives (goals) for adhering and for doing. We have to not just behaviorally conform, or adopt the goal and formulate the intention: We have to do that for specific "reasons" (recognition), for specific higher "motives": obedience, sense of duty, respect, and so on.

Sanction avoidance, the penalty, is not the *goal* that should motivate (drive) your adhesion and choice (the goal of the *sanction*—if any, apart from blame—is fairness, that you pay for your abuse, and that you and others be sent a message confirming N, its monitoring and equity: those who commit a wrong, who do harm (a public good like social order, or other people) have to pay.

"Sanctions" (penalties) are also for "learning" to pay "attention," to be aware of the existence of N and of its context and circumstances, and to not violate N out of unawareness. We have to become normative–attentive, and this is done by meta-Ns and imperatives ("Be responsible," "Be careful") and by shocks and learning.

3.2 Avoidance

As noted, Ns want to create a specific kind of goal, an *avoidance goal*. Or, better yet, they want us to perceive that imperative in an avoidance focus; it is a matter of

"framing": We have to frame the situation from the perspective of coercion, danger, duties, possible harms, and possible sanctions ("Prevention focus": Higgins 1997). This is because in this frame the decision is more coercive, and the prescription more effective.

In our view, the prospect of punishment and of sanctions, condemnation, violation, or being "bad" is precisely meant to place us in this cognitive and emotional "frame"; its purpose is more educational and communicative than practical and "economic," in part because we cannot detect and punish all deviations, and we need a strong *internal* control and precaution and later a feeling of guilt, as well as internal persecution and punishment.

The conflict between two avoidance goals, or between an avoidance goal and an approach goal, makes us feel a sense of astriction: We chose between a threat, a worry, and something else (negative or positive).

In fact, if the N has to eventually create or come into conflict with some other non-normative goal, it has to win. That is why the Ng takes the (psycho)logical form of an "obligation." Obligations are more stringent, cogent, this for two reasons:

(A) They are shaped in terms of "ought," that is, of necessity; that is, there are no alternatives, this is the only way: this or nothing (violation). It is not simply *useful* and within your choice; it is "imperative" in both senses. Even the technical "ought" is conceived as a necessity: "if you want x you have to y (if you cannot, you will not succeed)." In the deontic, normative "ought," this necessity perspective, this lack of alternative is even stronger, since—given that you don't know or understand or do not have to care about N's "end"—it is not up to you to see whether there may be some alternative way.

(B) If you do not do x, there will be a harm, a penalty, a bad situation for somebody and for you, something to be avoided.

This is why, in order to make the impinging goal more prescriptive and coercive, we make commitments and pacts even with ourselves. So we are bound, we are "in debt" and bound by duty to ourselves.

3.3 Meta-goals

N introduces a goal to do or not to do a given action, a goal that has to be "adhered to," that is, adopted because the source has the goal that you adopt it. So Ns are *meta-goals*: goals about a goal of yours, about a goal you *have to* adopt and pursue. And they want and have to win/prevail against your *possible* conflicting goals; there is a presupposition (like in any influencing action) of *possible* (and assumed) conflict with your autonomous goals; we do not suppose that you already and independently have that goal and would pursue it independently of N. However, this is not so crucial, since it makes a difference whether you do something for your personal motives or whether you engage in the same outward behavior but as an application of and respect for an N. The second case is markedly different, not by reason of possible sanctions

but in virtue of the real value and functioning of N. N is not a statistics of a given behavior; N is effective when that behavior is *olive to* the recognition of N as an N, to respect for the will of an authority, and you are also motivated by that (for moral principles *or* to avoid its sanctions).

So the goal of N is not just that you do something, but that you perceive/receive the N imperative, recognize it, and adopt it, and that you conform to it, and for that reason behave accordingly. Sometimes what matters is the subject's *obedience*, not the real content of the "order": to see whether he is submitted.

Ns do not necessarily entail a conflict in the subject. At any rate, the N usually creates some problem, being in conflict with other goals[14]: N can be in conflict with desires ("Do not covet your neighbor's wife"), with drives and needs ("Do not steal even if you are hungry"), with impulses ("Do not kill, even if you are furious"), or with other N-goals (conflict between Ns or their instantiations).[15]

Even in this case—involving N against other goals or one N against another N—the conflict can be of two kinds, or origins: logical contradiction (opposite goals) and resource scarcity (competing goals).

Moreover, in the decision-making process, the N-goal is subject to exactly the same scrutiny and steps as the other goals in what concerns the decision. For example, "I prefer to conform, but… are there the conditions and resources for that? Am I skilled, able? Are there possible side effects in this context to be avoided?"

3.4 Origin and Base of Norms: Norms Come from the Social Goals to Be Adopted

To better understand the meta-goal and adhesion-based nature of N, let us reflect on their forerunners. The origin and forerunner of Ns is not the frequency of a behavior or a hierarchy and authority, etc., but is a social goal: *my goal about your behavior*, and in particular (with socio-cognitive agents) *my goal about your goal*—a goal that impinges on you in that *you are expected to "adopt"* it (or, better yet, to "adhere" to my "request/prescription," a meta-goal).

The real origin of an *N as an N* is not just an expectation, a prediction (this is the origin of trust): It is *my social meta-goal* and our searching for some tool for influencing you, for inducing you to adopt that goal. The transition from mere expectation to (implicit or explicit) *prescription* is the step preparing "norms" in a deontic sense (Castelfranchi and Tummolini 2003; Castelfranchi et al. 2007). Real "authority" (and not just "force" based on fear or submission) is built on Ns, not the other way around.

[14] Or, better yet, the content goal derived from the N is in conflict with other goals I have.

[15] Notice that conflict is additional proof that an N is a goal and a goal source. In fact, conflict is a specific property of goals (of any kind: desires, drives, impulses, needs, projects, plans, intentions, values, etc.).

That is why *for an N be an N* (perceived and treated as such), I have to perceive the "prescription/request," not as your "personal" (private) request or goal, but as anonymous, impersonal, coming from the "collective" or from the representative or institutions of the collective, and it should not be addressed to *me*, as an individual, a specific person, but to a class of subjects, who may happen to consist of only one person, me; or to me as a role player.

4 The Relationship Between the Mental and External Goals of Ns

4.1 Norm Functions and Goals

As we said, frequently, the "functions" of an N are not fully understood (we may not be fully aware of them) and thus may not be intended (or at least may be "passively" intended or "accepted"): They may be unconsciously or unwillingly "pursued." Sometimes, there is even a paradoxical function (a kako-function, on which see note 2): a bad result, contrary to our subjective (individual or collective) goals, and even in contrast with the official objective of the N, but contributing to its reproduction and not just occasional or accidental.

There is also an important distinction here between two different kinds of Ns: ones that are explicitly issued, "deliberated" (by appropriate procedure and roles), like legal Ns or the official "rules" in an organization, versus conventional norms, emerging and established by tacit negotiation among participants. In the first case there is surely an intended result, a subjective goal of N. In *conventions*, this is less clear: Since nobody "decides" about N, and nobody necessarily decides its aim, or what it should guarantee, G is not explicitly in the minds of the agents.

Nevertheless, sometimes the G is understood (or supposed to be) and may also be approved, and we do not just obey N but "collaborate" with it. For example, the N prescribing that you do not stick our fingers in your nose in public has the G of not disturbing or disgusting others; we understand the meaning and aim of that N and respect it, or we respect it because we agree about that G or we do so in order not to be blamed, or simply for the sake of obedience and conformity with Ns, or both.

There is also a nonintended function of these kinds of Ns, which is to establish manners and rules of politeness, in such a way as to select people, give them a status and a membership, distinguish between levels in the population, preserve traditions, etc. Frequently, this function is really more important than the specific contents of politeness Ns, which can lack any sense, being just a ritual behavior.

An example of a nonintended (good) F of social Ns and customs[16] is the *reduction of uncertainty*, the right frame for reading events and activating the right (expected)

[16]These are not just frequent and regular behaviors but have a prescriptive component: People not only expect but want us to behave conformingly, and they critically react to any "violation" we may commit (Castelfranchi and Tummolini 2003).

behaviors (Garfinkel 1963); this reduction of subjective uncertainty, but also of cognitive and decision-making costs, and of negotiation costs, this "coordination" is an F of any social convention, habit, or script. Any community enjoys these benefits, and its customs, games, and scripts remain and are inherited and replicated for that benefit, too. But we do not *intend* that: We play that game for the specific results we obtain for our goals, not for maintaining the social order, trust, or uncertainty reduction.

Another F of any N—also in a strict sense (laws, moral Ns, social Ns, formulated in terms of an obligation or prohibition)—is *to teach and learn to "obey," as such*, any authority's will, any recognized N. It is a fundamental mental attitude, which makes us "social," acculturated: It is a fundamental cognitive and motivational basis of collective activity. We have to obey N and authority *as N* and *as authority*, not on condition that and because we understand and agree about N's goal. Our task is to "recognize" N and thus obey it, and by doing so, we send out a "signal," a message that that is an N, that we respect N and the authority, and that they should be respected. We do not intend that F of our behavior, but it is there and it reproduces and spreads it.

There are also *intended functions* of Ns; for example, we intend the (illusory?) effect of a stronger sanction for a given crime in order to dissuade people from that crime. N as the explicit goal of establishing that "if X commits that crime, s/he has to pay x amount of money, or has to serve a prison sentence of x years," and that these provisions be applied. But we also wish that an effect of this N be the reduction of that crime, and we change the law with this objective. This objective/G is not something that somebody "has to do," what is prescribed by N: The objective is some expected consequence—that is, the real aim of N. The punishment is increased only instrumentally in view of that outcome. Analogously, we establish Ns for giving right of way in traffic, etc., but we intend the general (expected) emerging effect: a regulation of traffic, reducing accidents and conflicts, a safer speed, making clear who is "in the right" in the event of an accident, etc. This is not what is prescribed (first-level goal) but is an *intended* F of it, and is also a G of it (at least in the mind of the issuer).

As noted there also are kako-functions: bad results that systematically reproduce N and a given behavior. What is (supposed to be) the goal and/or function of laws (and thus of legislators, of government)? (i) Social order and its maintenance or (ii) the "common good" and protection of the "commons." Imprimis (ii); (i) only if/since it is a "common good," but not if it favors a minority, some privilege, or class domination. Objectively, however, (i) the normative order (social, economic) is a subordination in favor of the dominant interests and groups. Ns are introduced or changed *in order to protect interests*: Every political, economic, or civil N protects some interest and subject. Who has the power to obtain the "right" N from the authority?

4.2 Subgoals

There is an obvious relationship between N's functional aim (F), its public effect/mission, and the G that Y (the subject) had to internalize and pursue (Gn): Gn is necessarily a subgoal for realizing F; the expected outcome of an action realizing Gn is F or part or a condition for it. Ns are aimed at planting in our mind—for regulating our behavior—a subgoal of the intended outcome (which is supposed to be a public good). For the formal (enounced, proclaimed) and official N (legal or not), there is usually a good overlap between N's intended goal and its F in the group/community, and we can even "adjust" N to its (ascribed) results. There is frequently a partial overlap, a partial explicitation of N's goal hierarchy: We issue N (and respect it) "for" G, which is a subgoal in the full goal "chain": goals for good Fs.

Society acts like evolution! On account of both our bounded rationality and limited computational resources and our personal preferences, we cannot understand and calculate the final, very high, and long-term and "complex" outcomes of our behaviors: their "fitness." This is why evolution and culture "terminalize" (as final motives) some subgoals of the real F, like in the relationship between evolutionary fitness and our internal "motivations," like in the creation of social "values" that had to be noninstrumentally justified as ends in themselves. For the same reason, Ns have to be obeyed (even) without understanding or sharing their aim. Society and culture reproduce the evolutionary trick.[17]

4.3 The Subject and N's Aim

As noted, N does not presuppose our understanding and sharing (pursuing) the *aim of N* (for the "subjects," it is just a presupposed F). But it does presuppose, or require, some belief that there is a goal, and more precisely a common end or value, and that the N authority (the issuer) is pursuing that goal. We have to *trust* the issuer and his playing a "tutoring" role.

We are "obliged" to obey even if we do not agree on N's goal. But we have to trust the issuer for his *intention* and role to do something "good," not self-interested. He may be wrong, but he should not be abusing of his power. If this is not the case, our impulse is not just to violate N but is stronger, because he is not playing his role, and his harming me/us. We feel entitled to rebel.

This "agreement" (common goal) about the Polis, Ns, authority, etc., is the background for issuing and respecting Ns. Some (not fully conscious) delegation, or

[17] N's other generic "function"—the restatement of a normative system, of authority, of submission—is also ensured by an internalized subgoal: the goal of adopting N (given its recognition as an N, which is another goal and subfunction of N).

empowerment (Gelati et al. 2004), and "alienation"[18] is intrinsic in a real normative process.

5 Concluding Remarks

Norms are artifacts, tools for manipulating human conduct through the manipulation of our goals and preferences/choices. It is impossible to understand the efficacy and working of norms without a modeling of *how Ns work in our mind, how they succeed in regulating our behavior from within*, and how do they give us goals. They are built for that.

However, Ns also have goals (they are aimed at achieving certain social outcomes), have effects (including unintended ones) and have functions. We do not understand and intend all the functions of Ns, and the subject is not supposed or requested to understand even all the goals of Ns and to obey on condition that she agrees and cooperates.

Let us conclude with a nice paradox of Ns.

N's goals/functions and their possible reversal. What is the real goal or function of a given N1? That we do not violate it, that our behavior conforms to N1; or that we be punished, as by paying a fine (norm N2)?

N2 (and its application) should in principle be only a means, a secondary and instrumental N designed to protect the realization of N1, that is, its goal. It is a meta-norm about the possible violation of another N, but functional to its *effectiveness*, not to its violation. But this *instrumental relation* can be reversed, and the means become the end, and the end merely a means, an excuse. I can issue N1 with the expectation (and goal!) that you violate it, so I can punish you. This is the attitude of some bad parents, but this is also, for example, the use of speed limits in some local government in Italy: They set an unreasonable speed limit *in order* to have a lot of people (not local citizens) violate it while traveling through their territory, *in order* to gain a hefty income from their fines. Hence, *N is issued in order for it to be violated!*

Even the subject can in some sense reverse the right goal relation, by interpreting and using the punishment (such as a fine) as just an additional cost for the possibility of exploiting the violation of N, and they decide to pay in order to be free to violate N1.[19]

[18] Meaning that the subject alienates his own intellectual evaluative, problem-solving, decision-making capabilities by "delegating" them to others, along with the power and the solution. Moreover, he is not in a condition to realize that, to understand this process, and behaves *without recognizing* his own alienated powers and without the possibility of reappropriating them. He has to be blind and to adopt N blindfolded.

[19] These are the famous findings of Uri Gneezy and Aldo Rustichini (2000).

References

Bargh, J., P. Gollwitzer, A. Lee-Chai, K. Barndollar, and R. Trötschel. 2001. The automated will: Non conscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology* 81: 1014–1027.

Bicchieri, C. 2006. *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Bratman, M. 1987. *Intention, plans, and practical reason*. Cambridge, Mass.: Harvard University Press.

Castelfranchi, C., and L. Tummolini. 2003. Positive and negative expectations and the deontic nature of social conventions. In *Proceedings of the 9th international conference on artificial intelligence and law*. Edinburgh: ACM.

Castelfranchi, C., F. Giardini, E. Lorini, and L. Tummolini. 2007. The prescriptive destiny of predictive attitudes: From expectations to norms via conventions. In *Agenti software e commercio elettronico: profili giuridici, tecnologici e psico-sociali*. Ed. Sartor, G., C.Cevenini, G. Quadri di Cardano. 43–55. Bologna, GEDIT.

Castelfranchi, C. 2012. Goals, the True Center of Cognition. In *The goals of cognition*, ed. F. Paglieri, L. Tummolini, R. Falcone, and M. Miceli. London: College Publications.

Castelfranchi, C. 2013. Cognitivizing norms. Norm internalization and processing. In *Law and computational social science*, ed. S. Faro and N. Lettieri. *Informatica e Diritto*, vol. XXII, 75–98.

Castelfranchi, C., and F. Paglieri. 2007. The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese* 155: 237–263.

Conte, R., and C. Castelfranchi. 1995. *Cognitive and social action*. London: UCL Press.

Conte, R., G. Andrighetto, and M. Campennì. 2010. Internalizing norms: A cognitive model of (Social) norms' internalization. *International Journal of Agent Technologies and Systems* 2: 63–72.

Elliot, A. 2006. The hierarchical model of approach-avoidance motivation. *Motivation and Emotion* 30: 111–116.

Garfinkel, H. 1963. A conception of, and experiments with, 'trust' as a condition of stable concerted actions. In *Motivation and social interaction*, ed. O.J. Harvey, 187–238. New York: The Ronald Press.

Gelati, J., A. Rotolo, G. Sartor, and G. Governatori. 2004. Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. *Artificial Intelligence and Law* 12: 53–81.

Gneezy, U., and A. Rustichini. 2000. A fine is a price. *Journal of Legal Studies* 29: 1–18.

Higgins, E.T. 1997. Beyond pleasure and pain. *American Psychologist* 52: 1280–1300.

Rao, A., and M. Georgeff. 1995. BDI-agents: From theory to practice. In *ICMAS-95*. In *Proceedings of the first international conference on multiagent systems,* ed. V. Lesser, 312–319. Menlo Park: AAAI Press.

Authority

Kenneth Einar Himma

1 Two Kinds of Authority: Epistemic and Practical

There are two kinds of authority: epistemic and practical authority. A person, *P*, has *epistemic authority* over a person, *Q*, with respect to the proposition *x* if and only if *P* sincerely asserting that *x* is true is, by itself, a reason for *Q* to believe that *x* is true.[1] An example of an epistemic authority with respect to cancer would be an oncologist. *P* has *practical authority* over *Q* with respect to the performance of act *a* if and only if *P* directing *Q* to perform *a* provides *Q* with a novel reason for doing *a.* Examples of practical authority include parents and judges.

Although practical authority differs from epistemic authority, it is worth noting, at the outset, two features common to each form of authority. First, each form of authority is characterized, at its conceptual foundation, as involving the capacity to create in another person *new* reasons of a theoretically significant kind. Second, in each case, the reasons that an authority's directive or statement provides to a subject are content-independent in the following sense: It is the source—and *not* the content—of the directive or statement that provides the subject with a reason.[2]

Nevertheless, practical and epistemic authorities are different in an important sense. Although both are properly characterized as authorities in virtue of their ability to give subjects new content-independent (or source-based) reasons, they are reasons of a different kind. An epistemic authority's opinion provides the subject with a novel reason to *believe* that opinion—provided that the content of the opinion falls within the scope of the authority's expertise. For example, a physician who tells me

[1] For a couple of notable discussions of epistemic authority, see Hurd (1999) and Zagzebski (2012).

[2] The notion of a content-independent reason will be explained in more detail below at pp. 9–10.

K. E. Himma (✉)
School of Law, University of Washington, Seattle, WA, USA
e-mail: himma@uw.edu

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_8

I have the flu has given me a new content-independent reason to believe I have the flu. In contrast, a practical authority's directive gives rise to reasons for *action*—in particular, a reason to do what the authority has directed the subject to do. For example, a judge who orders me to pay damages to *P* gives me a new content-independent reason to pay damages to *P*.

Indeed, it is commonly thought that legitimate (or morally justified) practical authorities have the capacity to bind subjects by providing reasons that give rise to a moral obligation to obey. It would be rather odd to think of epistemic authority as giving rise to reasons that *obligate* (as opposed to *oblige*) the subject to either believe what the epistemic authority opines or do what the epistemic authority recommends.[3] For this reason, an opinion from an epistemic authority should function cognitively in a different way from a directive by a practical authority.[4]

Again, while the notion of epistemic authority is relevant in addressing many issues, descriptive and normative, that arise in connection with the law, the concept of authority that figures most prominently in law is practical authority. Law is contrived, by nature, to issue directives that create new reasons for the subject to do what such directives require; after all, law, by nature, regulates behavior—and not thoughts. Accordingly, the remainder of this essay is centrally concerned with discussing practical authority.

2 Power, de Facto Authority, and Legitimate Authority

The most influential theorist on the nature of practical authority, Joseph Raz, distinguishes between de facto authority and legitimate authority (or authority per se). A legitimate authority is morally justified in issuing directives that tell subjects what to do and hence provide subjects with new reasons for doing what the authority directs them to do. In contrast, a de facto authority "either (1) claims to be legitimate or (2) is believed to be so" (Raz 1994, 211). As is readily evident, not every de facto authority is legitimate; some practical authorities might claim legitimacy without being legitimate, as is typically the case with totalitarian states. A de facto authority

[3] Although epistemologists have sometimes talked of the existence of epistemic obligations, it is not clear, given the seemingly involuntary character of belief formation, how a person could be *bound* to accept an epistemic authority's opinion on a matter within the authority's expertise; after all, we do not seem to choose what to believe. Further, while it is true that my consulting a physician, for example, acknowledges the physician's superior expertise on the matter, it is hard to see how that acknowledgment alone could *bind* me to accept her opinions—although her greater expertise surely gives me a reason to believe something.

[4] So much so that one might reasonably question whether the concept of "authority" applies in the epistemic context—if, as seems reasonable to think, the concept applies only to matters over which the subject of authority has direct volitional control. Of course, we are reason-responsive when it comes to belief, but that response is not determined by a free choice; that response is determined by a committed trust to the person's epistemic authority and is sometimes conditioned by our own sense of what is intuitively plausible, which is also not a matter of direct free choice.

is legitimate when either its claim that it has legitimate authority or its subjects' belief that it has legitimate authority is true.

Both de facto authority and legitimate authority should be distinguished from political *power*. A person can have power over another person without having authority. *P* can be an authority over *Q* only insofar as *P* is generally accepted as an authority in the relevant community (which might, or might not, include *Q*'s acceptance). All that is needed for *P* to have power over *Q* is that *P* has some reliable coercive means for inducing *Q* to comply with *P*'s commands. As is evident, the claim that *P* has power over *Q* is purely descriptive.

It is worth noting that, like the notion of power, the notion of de facto authority is purely descriptive while the notion of legitimate authority is normative.[5] As Raz puts the point, again, a de facto authority is someone who "either claims to be legitimate or is believed to be so, and is effective in imposing its will on many over whom it claims authority" (Raz 1994, 211). There are a number of ways one might conceive legitimacy: One might claim an authority is legitimate (1) when it has a "moral right to rule," as Raz sometimes puts it; (2) when its directives give rise to content-independent *moral* obligations to obey[6]; or (3) when its directives, in the case of coercive authorities, are justifiably backed by coercive enforcement mechanisms, from the standpoint of political morality.[7] But, however fleshed out, the notion of legitimacy is a morally normative notion insofar as it makes reference to someone's rights, obligations, or justified use of coercive enforcement mechanisms.

3 Practical Authority as Personal

The analysis of legitimacy has some interesting implications for the nature of practical authority. Insofar as legitimate authority has a moral right to rule or generates moral obligations to obey that are owed to the authority, it is a conceptual truth that practical authority is *personal* in the sense of being a subject of conscious experience. Only beings that are personal in this way can have a moral right to rule or can be owed moral obligations.

[5] For an extremely helpful discussion of the points sketched in this paragraph, see Christiano (2013).

[6] One interesting issue here concerns to whom the obligation to obey is owed. One could argue that the obligation is owed to the authority, which might facilitate justifying a coercive authority's use of enforcement mechanisms. Alternatively, one could argue that the obligation is owed to other subjects of the authority, as in, say, a fair-play argument that grounds an obligation to obey in one subject in the benefits that she receives from the compliance of others. But if Raz is correct in thinking that legitimate authority confers a moral right to rule on the part of the authority, then the corresponding moral obligations to obey are owed by subjects to the authority. Of course, this is compatible with the authority's generating obligations to obey that are owed to other subjects. Indeed, it seems quite plausible to think that subjects of morally legitimate authority have a moral obligation to obey that is owed to both subjects and the authority.

[7] These three conceptions of legitimacy do not necessarily coincide. As we will see, the differences between justifying coercive authority and justifying non-coercive authority are salient in giving adequate justifications for each. See Sect. 7, below.

Further, given the nature of practical authority and personal beings, an authority must be rational in the sense of being able to deliberate upon reasons in reaching a decision. For example, a lion might be able to direct a pride of lions, but it would be silly to characterize a lion as having practical authority over the other lions. Lions do not possess the relevant capacities to be capable of deliberations culminating in directives that give others reasons to act. This is not to affirm or deny that lions can have reasons to do things. It is merely to deny that lions can deliberatively issue directives in a sense that requires the sort of rationality that authorities must, as a conceptual matter, possess.

Similarly, a computing machine, given the technological limitations at this time, cannot have practical authority over a person, although it might well have epistemic authority over a person. A satellite navigation system might be an epistemic authority in that its output gives a person a new reason to believe that the system's directions are correct, but it seems implausible to think that it, by itself, gives rise to a new reason for action. Believing the directions are correct gives rise to a reason to comply with them, but that is different from claiming that the navigation system exercises practical authority. Of course, computing machines might someday have sufficient artificial intelligence capacities (which would presumably give rise to consciousness) that render them conceptually capable of being practical authorities. But those technologies are not sufficiently developed, at this juncture, to give rise to an example of artificial practical authority.

Likewise, an authority must be able to receive and issue communications to a subject regarding the authority's view about what the subject should do.[8] As the capacity for rationality is incorporated into the notion of authority, authority must be able to communicate in something that counts as a language. What properties something must have to count as a "language" is not entirely clear, but a language must surely contain semantic mappings from symbols to meanings and syntactic rules for articulating well-formed expressions in the language. Accordingly, it would seem that an authority must be capable of communicating in a language defined by semantic and syntactical rules.

If this is correct, this raises some puzzles about the logical connections between authority and authoritative directives. Any directive issued by an authority within the scope of her authority is clearly *authoritative*; thus, it is a sufficient condition that a directive issued by an authority within the scope of her authority is authoritative and provides subjects with a reason for action. But it is not obviously a necessary condition for a norm to be authoritative that it is directed or promulgated by an authority. If morality is objective and manufactured by a personal God, then the norms of morality are authoritative in virtue of being issued by a divine authority. Arguably, objective moral norms are no less plausibly characterized as "authoritative" if a personal God

[8] As Raz puts the point, "what cannot communicate with people cannot have authority over them" (Raz 1994, 217).

does not exist.[9] Either way, moral norms seem to provide subjects with the right kind of reasons for action.

More to the point, similar puzzles arise in connection with the authority of *law*. It is clear that individual valid laws—at least, those of a *legitimate* legal system—are directives that are authoritative over subjects who are within the jurisdiction of the relevant legal system.[10] What is not so clear, however, is whether the legal system that produces those authoritative directives should be considered as an authority.

The issue arises because a legal system, unlike a person's parents, is not *personal* in the sense usually thought requisite for being an authority. A legal system is neither an individual person nor some kind of compound person if there are such things. Rather, a legal system seems to be an abstract object that is constituted by various elements. Some of these elements are persons, such as individual executives, legislators, and judges; some of these elements are themselves abstract objects. Laws taking the form of rules, principles, or standards are normative propositions and are hence abstract objects.

It is important to understand that the defining property of abstract objects is that they are incapable of causally interacting with the world. There is nothing controversial about this among theorists working in the philosophy of abstract objects.[11] Consider, for example, the object denoted by the symbol "2". That object is the number for which "2" stands. Cursory reflection is sufficient to recognize that the number denoted by "2" cannot be seen, heard, touched, smelled, or tasted. In fact, the number denoted by "2" cannot be directly considered in thought; a symbol for that object is needed to represent that object in thought—although what that

[9] It is not clear whether Raz would characterize the norms of morality in such a case as "authoritative." One might deny, I suppose, that such norms are authoritative because they are not directives of an authority. I do not find that position plausible, but I cannot address the issue of whether there can be impersonal sources of authority—such as would be the case if moral norms are objective without having a personal being as author—in detail here. All I can do is raise the relevant concerns.

[10] As we have seen, a legal system might be believed to be authoritative, or claim authority, without actually being legitimate. It is not exactly clear whether, as a conceptual matter, the laws of an illegitimate legal system provide any kind of reason to act that is not connected with a prudential desire to avoid coercive enforcement mechanisms. One possible view is that such law merely "purports" to provide certain kinds of reason (usually thought to exclude the reasons provided by the authorization of coercive enforcement mechanisms), although what this would involve is somewhat unclear.

[11] As Gideon Rosen puts the point: "Concrete objects, whether mental or physical, have causal powers; numbers and functions and the rest make nothing happen. There is no such thing as causal commerce with the game of chess itself (as distinct from its concrete instances). And even if impure sets do in some sense exist in space, it is easy enough to believe that they make no *distinctive* causal contribution to what transpires. Peter and Paul may have effects individually. They may even have effects together that neither has on his own. But these joint effects are naturally construed as effects of two concrete objects acting jointly, or perhaps as effects of their mereological aggregate (itself a paradigm concretum), rather than as effects of some set-theoretic construction. Suppose Peter and Paul together tip a balance. If we entertain the possibility that this event is caused by a set, we shall have to ask which set caused it: the set containing just Peter and Paul? Some more elaborate construction based on them? Or is it perhaps the set containing the molecules that compose Peter and Paul? This proliferation of possible answers suggests that it was a mistake to credit sets with causal powers in the first place" (Rosen 2014).

representation amounts to, beyond its being ideational in character, is not clear. Thus, the propositions that constitute law are abstract objects that cannot causally interact in any way with subjects—although subjects can apprehend them through mental symbols that represent these abstract objects in a way that can be processed by rational beings in deliberating what to do.

Notice that a legal system is not an abstract object solely in virtue of including abstract objects, like the normative propositions expressed by sentences expressing laws; the legal system is an abstract object also in virtue of its being a collection of different types of object—or, otherwise put, a *set* of some kind. We cannot causally interact with sets if, as is commonly believed, sets are abstract objects. We can sometimes pick up members of a set if they are concrete individuals—such as, for example, an apple and an orange; what we cannot do is pick up the set made up of that apple and that orange with nothing more. If we put the apple and orange in a bag, we can, of course, pick up the bag; however, the bag containing the apple and orange is a concrete object that enables us to pick up the apple and orange together—and larger collections of fruits and other objects. But, strictly speaking, we are picking up the bag with the apple and orange, and not the set consisting of the apple and orange. Abstract objects cannot be picked up.

Like a set of any objects, a legal system is not something that can be touched, seen, smelled, tasted, or heard—although various elements of a legal system can be perceived in one of these ways. Indeed, like a set of fruits, a legal system is not the kind of thing that can be lifted and carried, and it is not because a legal system weighs too much. A legal system is simply not the kind of thing that can be lifted; not even an omnipotent God could lift either the number "2" or a legal system.

At first glance, this seems to create a problem for the view of a legal system as an abstract object. How could a law be authoritative if an authority does not issue or promulgate it? There are two potential answers here. First, as seen above, one could argue that it is not a necessary condition for a norm to be authoritative that some authority has promulgated it. Again, if morality is objective and not manufactured by God or some particular person, then the norms of morality can be thought of as authoritative despite not being issued by an authority.

Second, one could attempt to explain the authoritative quality of valid legal norms in terms of the authority of some particular official or perhaps the collective authority of *multiple* authorities, as opposed to a set of authorities. It is, of course, the cooperative work of these personal beings that culminates in the production of laws that are authoritative—and hence the authoritative quality of valid laws might well be inherited from the authority of these personal beings. If so, the authoritativeness of laws would be derivable from the actions of personal authorities in a way that is much more complicated than initially appears.

But it would be a mistake to try to suppress these complexities by attributing the requisite qualities of a personal being to something that is, as an uncontroversial conceptual matter, not capable in principle of doing what personal beings do. It is true that a legal system contains many objects that are capable of guiding a subject. Officials, as personal beings, can express a view about what subjects ought to do, and laws, as propositional objects, express content dictating what subjects ought to

do. Nevertheless, the legal system is constituted by the set of officials, norms, etc., and is hence as much an abstract object as any other. Accordingly, if a legal system is an abstract object and *if* only personal beings can be practical authorities, then a legal system cannot be an authority because it is not a personal being.[12]

4 Practical Authority and Its Reason-Giving Capacity

Legitimate practical authority is characterized by the ability to provide a subject with a *new* reason for doing what the authority directs the subject to do.[13] That is to say, a legitimate authoritative directive provides the subject with a reason for action that she does not have absent the directive (or norm). Thus, a legitimately authoritative directive provides a reason for doing what the directive requires that alters the subject's reasons for acting with respect to the relevant action.

This new reason is *moral* in character. As discussed above, to say that an authority is legitimate is to say, among other things, that the authority is morally justified in issuing directives that bind subjects. As such, the directives of a legitimate authority give rise to moral obligations. As moral obligations give rise to moral reasons, the new reason to which a legitimately authoritative directive gives rise is a moral reason.

It bears reiterating here that not all authorities are legitimate. A merely de facto authority does not have the capacity to give subjects a *moral* reason to do what the authority directs. A de facto authority might have sufficient coercive power that a subject has a *prudential* reason to do what the authority directs—in the form of a reason to avoid being subject to coercive sanctions. But a merely de facto authority (or illegitimate authority) has no general capacity to issue directives that *provide* subjects with a new moral reason to do what the authority directs.[14] At most, as the matter is commonly put, a de facto authority "purports" to provide moral reasons for action—although purporting seems, at first glance, to require certain personal communicative capacities lacking in abstract institutional authorities, such as law.

---

[12]This raises an interesting issue with respect to Razian positivism. Raz takes the position that it is a conceptual truth that law "claims" legitimate authority. On this view, while law's claim to legitimate authority can be, and often is, false, it is a conceptually necessary condition for a something to count as a legal system that it makes such a claim. If law is an abstract object of some kind, then it is conceptually impossible for law to make claims. See Himma (2001) and, for a very similar subsequent argument, Dworkin (2002).

[13]As Scott Hershovitz convincingly explains: "When one makes a request, one gives the addressee a reason for action that she did not have before … Countervailing reasons may outweigh the [reason provided by the] request. If I request that you help me carry my groceries, I expect you will consider my request along with all the other reasons you have for action. I expect you to act upon my request only if it tips the balance of reason in favor of doing so" (Hershovitz 2003, 204). Although he is expressly concerned with requests in this passage, the same considerations, as Hershovitz observes, apply to authoritative directives.

[14]Of course, subjects might have a moral reason to do what a de facto authority commands if the command reflects the requirements of morality, but that reason would not be a new reason that is explained by the authorities issuing the relevant directive. See, below, at p. 10.

Further, insofar as the directives of a legitimate authority are, as a general matter, morally justified, the new moral reason to which a particular directive gives rise is content-independent in the following sense: that a legitimate authority commands that I do φ gives me a moral reason to do φ regardless of what φ is. If "φ" stands for "cross the street," then the directive to φ gives me a moral reason to cross the street; if "φ" stands for "do not cross the street," then the directive gives me a moral reason not to cross the street. The directives of a legitimate authority give rise to content-independent moral reasons to do what the directive directs. Otherwise put, the directives of a legitimate authority give rise to *source-based* reasons that do not depend on the content of the directive.

An illegitimate authority's directives might also give rise to content-independent reasons for action, but they need not be moral in character. As noted above, an authority with *sufficient* capacity to coerce subject behavior might have a capacity, other things being equal, to give subjects a content-independent reason to act, but that reason will be prudential and not moral.[15] The reason will be grounded in the subject's prudential interest in avoiding being subject to coercive enforcement mechanisms, rather than necessarily in any content-independent moral reasons.[16]

One might think that an illegitimate authority can, under certain circumstances, issue directives that give rise to moral reasons for action. Insofar as there are moral reasons to follow a directive of an illegitimate authority, these reasons will be content-dependent in the sense that it is the moral content of the directive that provides the reason. For example, there is surely a moral reason to conform to the content of an illegitimate authority's directive not to kill innocent persons.

But to say that the authority's directive in this case gives rise to a *new* moral reason to act misrepresents the situation. The subject's moral reasons derive from the moral content of the directive and not from the fact that the authority has issued the directive. Since the directive of an illegitimate authority cannot give rise to the right kind of source-based reason to obey, it cannot give rise to a new moral reason to obey. What moral reason subjects might have to obey an illegitimate directive has nothing to do with the ostensible status of the directive's source as an authority.

Thus, while the directives of a legitimate authority give rise to content-independent moral reasons to act, an illegitimate authority, on Raz's view, merely "claims" (or "purports") that its directives give rise to content-independent moral reasons to act. An illegitimate authority's directives do not necessarily give rise to reasons that are either content-independent or moral. While it is clear that a morally illegitimate authority cannot issue directives that give subjects a content-independent *moral* reason to comply, it should also be clear that any illegitimate authority that lacks sufficient coercive ability or power over a subject lacks even the general capacity to give rise to content-independent *prudential* reasons for acting.

---

[15] As a merely prudential reason, this is a reason that can be outweighed by moral reasons.

[16] To say that a person is "subject" to coercive enforcement mechanisms is not to make any claim about the probability of incurring liability under such mechanisms. It is rather to say that the mere authorization of coercive enforcement mechanisms backing the directive gives a person some content-independent reason (though possibly quite weak) to comply with the directive.

5 Practical Authority and Its Capacity to *Bind* Subjects

The *new* moral reasons created by legitimately authoritative directives have a notable quality: Those reasons are sufficiently strong to create a *moral obligation* to comply with the authority—or so it is commonly thought. One of the most intuitively conspicuous features of authority is that its legitimate directives *bind* subjects. It is not merely a matter of a legitimate directive being something a subject *should* obey; it is rather a matter of a legitimate directive being something a subject *shall* or *must* obey. In some very difficult sense to specify (which does not in any way implicate the subject's capacity for free will), the subject of a legitimately authoritative directive is not "free" to disobey.

This is not limited to sources of authoritative directives that take the shape of a legal system. Any legitimate authority has the capacity to morally obligate subjects with a directive that falls within the legitimate scope of the authority. The purpose of an arbitrator, to take one of Raz's most influential examples, is to resolve a dispute between two persons regarding what should be done based on the reasons that antecedently apply to them. If the arbitrator may legitimately weigh the reasons and resolve the dispute *for the subjects*, then they are bound, morally (and possibly legally, depending on the facts of the particular case), to comply with the arbitrator's decision.

In practice, most instances of arbitration are reasonably thought to be legitimate because grounded in mutual promises of the parties to abide by the decision; even when ordered by a court, the directive is typically grounded in some kind of antecedent agreement between the parties. While the existence of a contractual obligation might not be a necessary element of the arbitrator–subject relationship, an exchange of promises to obey an arbitrator, regardless of what she decides, gives rise to at least a prima facie moral obligation to obey the arbitrator. But it is surely possible for arbitration that is imposed on the subjects to be legitimate and hence create moral obligations to obey in the relevant subjects.

In cases of coercive authority, the legitimacy of an authority has one very important implication: It would appear to morally justify the use of coercive enforcement mechanisms by the authority to ensure compliance. Insofar as coercion presumptively infringes upon a person's moral interests in acting autonomously (construed to include a moral interest in not being coercively required to perform, or abstain from, a particular act), it is morally problematic and requires some kind of moral justification. If a subject has a content-independent moral obligation to obey the authority's directives, then there is some reason to think that it is morally permissible for the authority to resort to coercive enforcement mechanisms as a response to non-compliance.

Of course, the issue is somewhat more complicated than that in both directions. The existence of a moral obligation to obey authority is not obviously a necessary condition for the authorities being morally justified in coercively imposing certain consequences for non-compliance. If a parent's power to direct his or her children includes a morally justified capacity to resort to (mildly) coercive sanctions to induce

compliance, that capacity cannot always be explained by the existence of a moral obligation to obey. Children do not come into the world with a developed capacity for moral agency that renders them morally accountable for their behavior; it is simply nonsense to think that, absent extraordinary circumstances, a three-year-old child has any moral obligations whatsoever.

Indeed, the example of parental authority creates a problem for the common view that the reasons created by a legitimately authoritative directive are moral in character. Although it is probably true that there are degrees of moral agency and moral accountability and that children become full moral agents by gradually acquiring the properties giving rise to moral agency, there will still be some children who seem subject to legitimate parental authority without any degree of moral agency (e.g., a two-year-old child). While other older children will be increasingly subject to acting according to moral reasons over time, very young children, fully lacking the relevant capacity, will not be subject to moral reasons. Indeed, it is reasonable to think that—if the language of reasons does not apply to very young children—such children are likely to comply out of some sense of prudential interest. At the earliest stage, rearing a child is a matter of developing certain stimulus-response mechanisms that will eventually culminate in views that will become moralized through the socialization process.[17]

Moral agency requires both the capacity for free action and a capacity for rationality that is sufficiently developed to support some threshold level of understanding of core moral requirements. Children may come into the world able to *choose* freely—and that much is incorrect if, as seems reasonable, free choice requires the ability to rationally weigh reasons—but they clearly do not come into the world with a sufficiently developed capacity to understand core moral requirements that would warrant either imputing obligations to them or holding them accountable for breaching such obligations. If this is correct, then authority can be morally justified in imposing coercive consequences for non-compliance on a subject without her having a moral obligation to comply.[18] Thus, the existence of a content-independent moral obligation to obey on the part of a subject is not a necessary condition for an authorities being morally justified in coercively enforcing a directive.

Nor is the existence of a content-independent moral obligation to obey an authority's directives a sufficient condition for a justified application of coercive enforcement mechanisms. Inducing compliance in a subject by coercive means remains

[17] Indeed, this description conforms to the first level of moral development in the theories of both Lawrence Kohlberg and Carol Gilligan. See Kohlberg (1984) and Gilligan (1982).

[18] In the case of parenting, these coercive consequences are not properly characterized as being "punishment" in any sense that includes a retributivist notion that the relevant unpleasant consequences are, as a moral matter, *deserved*. Parental discipline of young children can be characterized as "punishment" in a less robust sense that does not involve moral connotations of the disciplinary actions being deserved by the child. Parental discipline might, of course, be morally warranted by the parent's moral duties to rear a child to have certain character traits and behave in certain ways. But "punishment," in the robust sense of the word, connotes that the unpleasant consequences are morally deserved by the non-complying behavior of the subjects. As can be seen, the issues that arise with respect to authority and coercive enforcement mechanisms are quite complex—and, for that reason, cannot be addressed in more detail here.

morally suspect in the sense that we lack a theory that shows clearly that authorities are morally justified in enforcing their directives. If the content-independent moral obligation is owed to someone other than the authority, then the authority is not the one who is wronged by non-compliance. While merely being wronged by an act does not necessarily justify imposing coercive sanctions or inducements to act, it is not a trivial matter to see how someone could be justified in imposing such measures for non-compliance if non-compliance does not result in a wrong to the person seeking to impose such measures. While the law frequently allows for such practices, the question is whether and how those practices are justified.

Of course, there are other elements that might be present in the authority–subject relationship that could justify the use of coercion. In the case of a private arbitrator, the subjects might contract with each other to abide by the arbitrator's decision subject to certain coercive penalties. In this case, the arbitrator does not seem fairly characterized as having been wronged by a non-complying party and yet is justified in imposing the penalty. What does the necessary moral work in this case is the mutual exchange of promises supported by consideration (i.e., the contract); these features give rise to morally protected reliance interests that might justify coercive enforcement mechanisms. The existence of a content-independent moral obligation to obey the authority does not suffice, by itself, to morally justify the authority's imposition of coercive penalties for non-compliance.

6 The Kind of Reasons to Which Legitimate Directives Give Rise

As we have seen, legitimately authoritative directives are commonly thought to create content-independent moral obligations to obey and hence provide some type of special reason for action.[19] That an act is morally good provides a reason to perform that act, but it is not a reason that *binds* the subject. Moral obligations bind subjects and hence provide reasons for action that are considerably more robust in the sense that the subject of an obligation does not have an option not to comply; a subject, in contrast, should, but need not, perform an action that is morally good.

Moral obligations can thus be seen as providing a reason that is "final" in the sense that it either *is not* or, depending on one's metaethical view, *cannot* be defeated by other reasons.[20] Each possibility requires a different formulation. Raz's notion of a

[19] See the discussion on parental authority and the moral incapacities of young children, above, at p. 11.

[20] According to William Frankena, morality is "supremely authoritative"; on this plausible view, moral obligations claim supremacy over all other obligations—including legal: When moral obligations come into conflict with other obligations and practical considerations, the moral obligations win; the only thing that can defeat a moral obligation is another more important moral obligation (Frankena 1966, 688–696). Similarly, Bernard Gert describes this feature of morality as follows: "Among those who use 'morality' normatively, all hold that 'morality' refers to a code of conduct that applies to all who can understand it and can govern their behavior by it. In the normative sense,

conclusive reason captures the weaker idea that what I have called a final reason is one that is not, as a contingent matter, defeated by other reasons. As Raz defines the notion, "*p* is a conclusive reason for *x* to ϕ if, and only if, *p* is a reason for *x* to ϕ (which has not been cancelled) and there is no *q* such that *q* overrides *p*" (Raz 1975). It is crucial to note that, according to Raz's definition, a conclusive reason is one that, as a matter of *contingent* fact, *is not* outweighed or overridden by other countervailing reasons. That is, a merely conclusive reason would be final with respect to outweighing any other reasons obtaining in *the* actual (and thereby in one specified contingent) world.

In contrast, Raz's notion of an absolute reason captures the stronger idea that a final reason is one that *cannot* be defeated by other reasons; that is, that there is no logically possible world in which an absolute reason is outweighed or overridden by countervailing reasons. As Raz defines this notion, "*p* is an absolute reason for x to ϕ, if and only if, there *cannot* be a fact which would override it; that is to say, for all *q* it is never the case that when *q*, *q* overrides *p* (Raz 1975, 27; emphasis added)."[21]

This much about Raz's view seems uncontroversial. Morality is thought to trump all other considerations in the following sense: Only a more important moral obligation can provide a reason for action that defeats the reasons for action provided by a less important moral obligation. Accordingly, if *P* has a moral obligation to do φ, then *P* is bound to do φ, regardless of other considerations—prudential or otherwise. Morality is supreme in terms of the force of the reasons moral obligations provide. Insofar as this is so, it follows that the reasons morality provides are final in the sense described above.

Accordingly, each of the types of reason Raz defines is a potentially accurate description of the final reasons morality provides depending on whether or not morality is objective or conventional in character. If conventional in character, then the truth-value of any moral principle is contingent, since it depends on the contingent views or practices of those who determine the content of the relevant convention; in this case, the appropriate sense in which reasons are final would be that they are *conclusive* in character. If objective in character, then a moral principle is necessarily true, if true at all[22]; in this case, the appropriate sense in which reasons are final would be that they are *absolute* in character.[23]

One of the most influential features of Raz's theory of practical reasoning is an account of the nature of the final reasons that apply in morality and of the way in which they either characteristically or should function in moral deliberations. Raz

---

morality should never be overridden, that is, no one should ever violate a moral prohibition or requirement for nonmoral considerations" (Gert 2012).

[21]There is an ambiguity in Raz's formulation of the two kinds of reason. While the first clause uses the modality "cannot," the second employs only a variation of a universal quantifier ("never"), suggesting that there is, in fact, no overriding *q*. To avoid replicating the notion of a conclusive reason, the notion of absolute reason should be construed as intending the modal clause. Otherwise, there is little difference between the two concepts.

[22]For example, on an objectivist view, it *cannot* be morally permissible to torture infants for fun.

[23]One can, of course, disagree that moral objectivism implies that moral judgments are necessarily true, if true at all. If so, then Raz's notion of a conclusive reason would apply to an objective morality.

calls this specific type of reason a *protected reason*, which consists of a first-order reason (i.e., a reason having to do with actions or beliefs) to do what the authoritative directive or valid moral norm requires, together with a second-order reason (i.e., a reason for acting, or not acting, on some set of first-order reasons), which he calls an *exclusionary reason*.[24]

As the notion of a first-order reason to do what a norm or directive prescribes is straightforward, the remainder of this section will be concerned with explicating the notion of an exclusionary reason. To begin, note that the notion of an exclusionary reason is compatible with each conception of morality and each of Raz's conceptions of a final reason. A reason, for example, could be exclusionary in all logically possible worlds, or it could be exclusionary in some worlds, but not others, depending on the specific circumstances of the possible world.

According to Raz, an "exclusionary reason is a second-order reason to refrain from acting for some reason" (Raz 1975, 39). Normally, practical rationality requires of a person that she weighs all of the relevant reasons and acts on her assessment of the balance of reasons.[25] Exclusionary reasons operate to exclude certain reasons from the reasons on which a person can rationally act. They do not prevent her from deliberating to determine what the balance of excluded reasons would require by way of acting; they simply preclude her acting on the basis of those reasons or her assessment of those reasons.

On this view, for example, a moral obligation not to kill an innocent person provides an exclusionary reason that precludes the agent's acting on the basis of certain reasons she might take herself to have to kill an innocent person, such as a desire to kill the person for the purpose of taking her belongings. This element of

[24] Hershovitz does a characteristically elegant job of explaining the notion of a second-order reason: "What does it mean to have a second-order reason, a reason to act for or not act for another reason? An illustration will help. Suppose Aaron's grandmother is in the hospital and that this provides Aaron reason to visit her. Suppose further that Aaron goes to the hospital and visits his grandmother, but only because he was hoping to run into Michelle, whom he has a crush on. In this case, Aaron conforms to his reason to go to the hospital to visit his grandmother but he does not comply with it. Does Aaron have reason to comply with his reason to visit his grandmother rather than just conform with it? Raz suggests he does, and I agree. Because Aaron went to the hospital to see Michelle and not his grandmother, his actions do not embody appropriate respect for his grandmother. Aaron had a reason to comply with his reason to visit his grandmother, that is, he had a second-order reason to act for a reason: Only through visiting his grandmother for the sake of visiting her could he show her proper respect" (Hershovitz 2003, 202).

[25] Heidi M. Hurd, for example, argues it can never be practically rational to accept exclusionary authority because it violates the principle that an agent should always act on the balance of reasons available to her (Hurd 1991). As Thomas May makes the point: "Acting on what the authority judges ought to be done appears to circumvent one's own evaluational judgement, and thus autonomy. By circumventing the evaluational judgement of the subject it seems the subject is *prevented* from acting on her own determination of what ought to be done. The subject seems to be eliminated from the determination of her behavior" (May 1998, 130). It is worth noting that even if it is practically irrational to accept authority in the sense of providing exclusionary reasons, it does not follow that authority is necessarily illegitimate. This could simply be taken to imply that consent would be no part of a successful theory of state legitimacy and that other moral considerations would be sufficient.

exclusionary reasons is primarily negative in character: It simply precludes acting on the basis of some specified set of reasons.

Of course, a moral obligation not to kill also provides a *new first-order* reason not to kill, which makes it another example of a protected reason, but its exclusionary character distinguishes it from other reasons having to do with whether to kill in the following respect: It is not a reason to be weighed in the balance with other reasons for or against killing; it is a first-order reason not to kill coupled with a second-order reason not to act on a specified class of reasons that would include, for example, any prudential reasons.

On Raz's view, a legitimately authoritative directive requiring P to do φ creates a first-order reason for P to do φ and a second-order exclusionary reason not to act on a specified class of other first-order reasons. Again, exclusionary reasons do not preclude an agent from deliberating to determine what the balance of excluded reasons require; they simply preclude the agent from acting on those reasons or on her assessment of what those reasons require.

In contrast, H. L. A. Hart took the position that legitimately authoritative directives provide *peremptory* reasons that preclude the agent from even deliberating on the class of excluded reasons (Hart 1982, 253).[26] On this view, an agent who has a peremptory reason to do φ would be violating norms of practical rationality (which presumably incorporates norms of both morality and prudence) simply by deliberating on—or weighing for herself—the class of excluded reasons.

Hart's view is problematic. Neither the norms of morality nor the norms of practical rationality are directly concerned with an agent deliberating on reasons. Morality is chiefly, if not exclusively concerned, with what an agent does in the world—not with how an agent deliberates on reasons that do not ultimately culminate in her performing some act.[27] Practical rationality is concerned with evaluating the rationality of a person's actions, and this is done by assessing whether the agent's acts conform to the balance of what Raz sometimes calls "right reason."

There might be exceptional cases in which norms of rationality or morality would condemn an agent for merely having a thought or considering a reason—even if she does not act on that reason. Consider, for example, a person P who consciously harbors a white supremacist worldview that is grounded in the thought that persons of other races are "subhuman" and undeserving of moral respect. Suppose P never acts either on those views or in any way that would express those white supremacist

[26] As Hart puts the point: "[T]he commander characteristically intends his hearer to take the commander's will instead of his own as a guide to action and so to take it in place of any deliberation or reasoning of his own: the expression of a commander's will that an act be done is intended to preclude or cut off any independent deliberation by the hearer of the merits pro and con of doing the act…. This, I think, is what is meant by speaking of a command as 'requiring' action and calling a command a 'peremptory' form of address" (ibid., 253).

[27] It is true, of course, that the moral evaluation of a person's action will also include consideration of her motive for acting. Giving to charity for the reason that it will help the poor, for example, is a morally valid reason to give to charity; doing so to enhance one's reputation in the community is not. But a person's motive has to do only with the actual reason on which she acted and not on how she arrived at that reason. There might be instances in which the reasoning comes into play, but this is not how one's mental states are characteristically evaluated from a moral point of view.

views in a harmful fashion. Indeed, assume that *P*'s outward behavior evinces equal treatment and respect for all persons—white and nonwhite—and that *P*'s outward behavior is so benign when it comes to differentiating persons on the basis of race that no one would ever think to characterize *P* as a racist.

It is reasonable to think that the mere holding of such views is morally culpable, but this is not an unproblematic position. Again, morality seems characteristically concerned with outward behavior and not inner mental states, in part, because it is not clear to what extent the relevant inner states—our beliefs and our desires—are within a person's volitional control. It is plausible to think that some mental states are subject to moral evaluation; hate might be one such example, along with the racist views considered above. But if these are subject to moral evaluation, these states would seem to constitute either borderline or otherwise exceptional cases.

If it is somewhat unclear whether morality or practical rationality applies to the mental states and acts discussed above, there is little reason to think that either precludes merely deliberating on a set of excluded reasons. Perhaps norms of morality and practical rationality would condemn being tempted to act on one's deliberation on excluded reasons after one has arrived at a view about the balance of those reasons, but it seems implausible to think that those norms would condemn deliberating on excluded reasons out of curiosity or out of a desire to see whether the directive lines up with the outcome of one's own deliberations. If so, Hart's view that legitimate directives give rise to peremptory reasons is false.

Raz's view is, clearly, not subject to such objections. Raz's view harmonizes much more closely with morality's significantly greater concern with actions than with thoughts. Raz's account of an exclusionary reason does not preclude deliberating on the balance of applicable and excluded reasons; it merely precludes acting on the basis of those reasons.

This certainly conforms to our intuitions about authority in general. If an arbitrator legitimately decides a contested issue between *P* and *Q* with a directive that binds them, there seems nothing either practically irrational or morally problematic with either *P* or *Q* deliberating on the excluded reasons. Certainly, neither the parties nor the arbitrator seems to have any grounds for complaint or criticism in such a case.

Nevertheless, Raz's view is subject to some questions and concerns. To begin, there are other mechanisms than that of an exclusionary reason for capturing the idea that authoritative directives provide reasons that are final in the relevant sense. As Stephen Perry has persuasively observed, an authoritative directive need not function as a second-order reason to do the conceptual work authority is thought to do (Perry 1989). If sufficient weight is assigned to the directives of authority, it might simply outweigh all the other applicable reasons in the vast majority of cases—which would be enough for authority to be fairly characterized as capable of performing its conceptual function of telling people what to do by issuing directives making certain behaviors mandatory.[28]

[28]This is a view of authoritative reasons that would not run afoul of Hurd's view that it can never be rational for a person to accept *exclusionary* authority. Authority, as Perry conceives it, is not exclusionary, as it provides only strongly weighted reasons, rather than exclusionary reasons.

Further, it is not clear that Raz's account correctly applies to all forms of authority. One particularly salient form of authority that does not seem to cohere to the Razian account is law. It is clear that, on Raz's view, the valid legal norms of a *morally legitimate* legal system—i.e., a legal system that is an "authority," rather than just a "de facto authority"—would give rise to exclusionary reasons for action; morality seems to be a paradigm for systems of norms that give rise to exclusionary reasons in the sense that if any system gives rise to such reasons, morality does. What is not as clear is whether a legal system that has purely de facto authority—i.e., one that is not morally legitimate—gives rise to exclusionary reasons.

The question, then, is whether it is a conceptual truth that law gives rise to exclusionary reasons or, otherwise put, whether law as such gives rise to such reasons, which would entail that even illegitimate legal systems provide such reasons. Raz seems to answer the question in the affirmative:

> The legal point of view and the point of view of any other institutional system is an exclusionary point of view. Legal norms may conflict and in deciding what, according to law, ought to be done one may have to balance different conflicting legal considerations, but law is an exclusionary system and it excludes the application of extra-legal systems (Raz 1975, 145).

Likewise, Raz states that "an authoritative determination of a primary organ to the effect that *x* has a duty to perform a certain action is an exclusionary reason for *x* to perform that action" (Raz 1975, 145).

One issue that arises in connection with Raz's view here is what the source of the exclusionary reason would be. On Raz's view, prudential concerns to avoid being subject to coercive enforcement mechanisms define reasons for action but the reasons are "of the wrong kind" (Raz 1975, 145); other things being equal, prudential concerns are first order in character. Sanctions, thus, define, as Raz puts it, "auxiliary" reasons and not exclusionary reasons (Raz 1975, 145).

On Raz's view, it is a conceptual truth that law provides reasons that are unrelated to either the authorization of coercive enforcement mechanisms or morality. On this view, "it is the fact that those actions are required by law … [that] is the reason for performing them" (Raz 1975, 155). Otherwise put, the claim seems to be that one should comply with the law because it is law.

At first glance, if this is Raz's view, it is puzzling. It is hard to understand—and especially under a positivist view of the sort Raz holds—how law *as such* could give rise to reasons of any kind not related to the possibility of incurring sanctions. From the standpoint of practical rationality, it is hard to see why the mere fact that a norm was promulgated according to a social rule of recognition (i.e., has a social "source") would give rise to a reason of any kind. Once sanctions are subtracted from the picture, it is not clear what features of law would do the necessary reason-giving work.

As it turns out, Raz seems to take a weaker and more plausible position. On his view, law *as such* provides reasons only to those who have accepted the law and thus take what he calls "the legal point of view." As he describes the notion:

> The ideal law-abiding citizen is the man who acts from the legal point of view. He does not merely conform to law. He follows legal norms and legally recognized norms as norms and accepts them also as exclusionary reasons for disregarding those conflicting reasons which they exclude (Raz 1975, 171).

Further, he holds that it is not conceptually necessary for the existence of a legal system that citizens take the legal point of view—or even that they, from the standpoint of morality, should take the legal point of view; as Raz puts the point, "[i]t is not necessary for a legal system to be in force that its norms subjects are ideal law-abiding citizens or that they should be so (i.e. that legal norms are morally valid)" (Raz 1975, 171). Rather, he holds "it is necessary that its judges, *when acting as judges*, should on the whole be acting according to the legal point of view" (Raz 1975, 171).

It should be noted that the above quote calls attention to an ambiguity in the claim that law *as such* provides exclusionary reasons. On one interpretation, the claim states that every valid *legal norm* provides an exclusionary reason. On another interpretation, the claim states that every *legal system* provides some exclusionary reasons. The above quote suggests that Raz has in mind the second interpretation, which makes a much weaker conceptual claim about law than the first. Here, it is important to note that the second quoted sentence in the last paragraph states only that judges "should on the whole" regard the law they apply as exclusionary reasons; it is not conceptually necessary that they regard every such law as exclusionary.

Raz's view, as expressed above, is weaker in a second sense. According to the above quote, what is conceptually necessary to the existence of a legal system is not that it provides exclusionary reasons for citizens. Rather, what is essential is that law provides exclusionary reasons to a particular subclass of *officials*—namely judges. What is notable here is that, on Raz's view, judges must *accept and treat* not only duty-creating recognition norms as exclusionary, but also the first-order norms they apply to subjects.

As regards citizens, then, one might take the view that it is a conceptual necessity that law as such merely "purports" to provide citizens with exclusionary reasons insofar as law is enforced in an exclusionary manner. In contrast, if morally legitimate legal systems give rise to content-independent moral obligations, then the laws of a morally legitimate legal system would provide citizens with exclusionary reasons that would be moral, and not legal, in character.

Although Raz's claim is frequently understood to be that law provides citizens with exclusionary reasons, there is good reason to reject that claim, as Raz appears to. There are many valid laws of legitimate legal systems that do not seem plausibly characterized as giving *citizens* exclusionary reasons. For example, I habitually park illegally because it is profitable to do so. The law prohibits parking without paying a fee and imposes a fine for non-compliance. The fine is $45, compared to the cost of parking, $10. I park illegally five times a week and get a ticket, at most, once a month. Assuming a month has four weeks, my net saving is $155—a profit that, on my prudential calculations, makes parking illegally the rational thing to do.

Illegal parking seems neither necessarily morally wrong nor necessarily practically irrational, but it can be problematic from either standpoint in certain cases. If one illegally parks in front of a fire hydrant, one is creating a risk of harm to

others—and that seems problematic from the standpoint of both morality and practical rationalities (conceived of, as Raz's account does, as concerned to identify what right reason requires). Similarly, if one takes up a parking space for significantly longer than the time allowed for that space, one arguably acts in a way that is problematic from each vantage point. Finally, if one parks illegally in a space reserved for disabled persons, then one is acting in a morally problematic way.

But more mundane instances seem consistent with both morality and practical rationality. If I park illegally for half the time that anyone is allowed to park in the space, it is far from obvious that I have done any moral wrong or violated norms of practical rationality. Parking laws are enforced by comparatively minor fines, and there is little stigma, moral or otherwise, attached to such mundane instances of illegal parking.[29]

Nor is it any more plausible to think that officials regard such laws as giving rise to either moral obligations or exclusionary reasons. While the model of conceiving certain laws as simply defining costs for non-compliance surely cannot be generalized across all areas of law (e.g., criminal law), it seems eminently plausible in the case of "municipal offenses." Indeed, I have frequently come back to my car just as a police officer was about to write a ticket and persuaded her to let me go. When an officer does let me go, she does so with just a lecture and warning—suggesting that, at some level, she understands and *condones* my reasoning.

This is reason enough to reject, as Raz seems to, the idea that law must provide citizens with exclusionary reasons, but there are other laws than those defining municipal offenses that are not plausibly regarded as giving citizens exclusionary reasons to comply. Consider, for example, contract law, and suppose that P and Q make a contract for P to do φ. Suppose that it is far more profitable for P not to do φ and to pay what a court requires as damages for breach. Suppose, further, that Q is indifferent with respect to whether P performs or pays court-ordered damages. Finally, suppose that P and Q know all the facts and that Q will choose to litigate if P breaches, rather than to settle.

From the standpoint of practical reason, the rational thing for P to do—from each party's vantage point—seems to be to breach and pay damages. Since Q is not made worse off by the breach and P is made better off by the breach, it would seem rational for P to breach, unless there is some *moral* reason requiring P to perform, a claim that seems implausible. However, the reasons adduced above seem to exhaust the class of relevant reasons; it seems silly to claim that an overriding reason would be "because the law says so." Thus, the rational thing to do seems to be for P to breach—and, notably, for reasons that are *prudential* in character.

Looking even at official practice, the rational thing for P to do seems to be to breach. From the standpoint of the legal system, there is nothing that would seem to entail that officials expect the parties to perform under the contract. If officials conceived of contract law as providing exclusionary reasons, then the appropriate legal remedy for breach would be specific performance. But courts in every

[29] Similar claims can be made about crossing against a red light in the middle of the night when no one is on either of the roads intersecting at a crosswalk.

jurisdiction of the USA prefer ordering payment of money damages to ordering specific performance. A special showing that money damages cannot adequately compensate for the breach is a legal requirement for ordering specific performance. But if a special showing is required to justify a court order of specific performance of the very thing *P* contracted to do, then the law itself seems not to view the mandatory norms requiring mutual performance of contractual duties as giving rise to exclusionary reasons.

Indeed, it seems implausible to think that anyone *should* have any other attitude toward the violations of such laws. It is surely reasonable to think that criminal laws provide exclusionary reasons—if any do—but the whole point of cutting the law up into the different areas of criminal and civil, with the different enforcement mechanisms, suggests that officials themselves believe that different types of offenses give rise to different levels of normative force.

It is true, of course, that valid legal norms—even those of the civil law—tend to be enforced in a way that excludes certain legal excuses or justifications, but this does not imply that the law provides, or even purports to provide, exclusionary reasons. Courts typically, though not universally, enforce some kind of remedy for violations of the law that are motivated only by prudential reasoning—suggesting that the law "excludes" certain justifications for non-complying behavior as excuses that will preclude application of the law's coercive enforcement mechanisms. Even so, the fact that law is enforced in such a way entails nothing as to the structure and functioning of the reasons law provides. In particular, exclusionary enforcement of law does not imply either that subjects characteristically do, or should, regard the law as providing exclusionary reasons.

7 The Justification of Practical Authority

To the extent that the exercise of practical authority appears to infringe autonomy rights, it raises a number of normative issues. First, insofar as the exercise of practical authority seems to infringe on autonomy rights, it raises a moral issue: What conditions must practical authority satisfy to be morally justified? Second, insofar as it is inconsistent with the principle requiring that one acts on the balance of reasons, it raises an issue of normative practical rationality: Under what circumstances is it practically rational to accept an authority and act on its directives?[30]

The importance of justifying practical authority—especially state authority—arises because there are two different forms of philosophical anarchism challenging the idea that many states are legitimate. According to *strong philosophical anarchism*, no state can be legitimate unless subjects rationally consent to its authority; however, it is not rational for subjects to consent because people have a moral duty not to allow an authority to substitute her judgment for theirs in deciding what to do (Wolff 1970). According to *weak philosophical anarchism*, it is possible for a state to

[30] See, e.g., Hurd (1991), and note 25.

be legitimate on the basis of subject consent because people have a moral *right*, which can be waived by consent, not to be bound by the state's commands—rather than a moral *duty* not to allow the state to preempt their own judgment (Green 1989; Simmonds 2001).

Complicating the task of justifying legitimate authority is that practical authority can take different forms that have morally salient different features. Some practical authorities lack the capacity to coercively enforce their directives. For example, one might think a physician is a *practical* authority in the areas of her expertise (i.e., *epistemic* authority) and can issue what appears to be authoritative directives (e.g., "take two aspirins and call me in the morning"), but a physician lacks any capacity to coercively enforce her directives.[31] In contrast, many practical authorities have the capacity to coercively enforce their directives. For example, an arbitrator to a dispute typically has some coercive mechanism at her disposal to enforce her resolution of the dispute.

Of course, the most important example of an authority with the capacity to coercively enforce its directives is a legal system. Regardless of whether or not the authorization of coercive enforcement mechanisms is a conceptually necessary feature of a legal system,[32] every existing legal system of which we know is authorized with the capacity to coercively enforce the law—and does so, sometimes in a ruthless and discriminatory way. Tragically, this is happening in the USA with increasingly frequent and unjustified shootings of unarmed black persons by police with little apparent appreciation of the fact that black lives matter.

This feature of law gives rise to what is the defining problem of normative political philosophy. The problem is how to justify the state's doing what no one else is permitted to do—namely issue commands backed by the threat of violence. In this respect, the state resembles an armed robber whose demand for the victim's money is backed by the threat of force. There are, of course, many philosophical attempts to state the conditions that determine when the state is morally justified in using coercive enforcement mechanisms to induce compliance with the law or punish noncompliance—the theories of John Rawls and Robert Nozick being two of the most highly influential in contemporary political philosophy.

As interesting as these theories are, they are not relevant here, as they would not apply to all forms of practical authority. Neither Rawls's theory of justice (i.e., the principles of justice he believes would be selected from the original position) nor Nozick's libertarianism is even remotely relevant with respect to the normative issues involved in accepting and complying with a physician's practical authority—or, if

[31]The issue of whether a physician has practical authority is a difficult one, but not much turns on it here if, as many theorists believe, there can be practical authorities that lack the authorization to coercively enforce directives. Joseph Raz, for example, argues that law—and hence legal authority—would be needed to resolve certain disputes in a society of angels who are always conclusively motivated to obey the norms that resolve those disputes (Raz 1975, 159–160; Shapiro 2011, 169–170). For a response to this argument, see Himma (2016).

[32]I argue that the authorization of such mechanisms is a conceptually necessary feature of law. See Himma (2017, 593–626).

one is skeptical that physicians have practical authority over a patient, the practical authority of any agent that lacks authorization to coerce compliance.[33]

In contrast, Joseph Raz has a more comprehensive normative theory that is intended to state the conditions under which practical authority is justified—one that would cover the practical authority of someone who lacks the ability or authorization to coerce compliance. According to the normal justification thesis ("NJT"), authority is justified to the extent that the subject is more likely to do what right reason requires by following authoritative directives than by following her own judgment:

> The normal and primary way to establish that a person should be acknowledged to have authority over another person involves showing that the alleged subject is likely better to comply with reasons which [objectively] apply to him (other than the alleged authoritative directives) if he accepts the directives of the alleged authority as authoritatively binding, and tries to follow them, than if he tries to follow the reasons which apply to him directly (Raz 1994, 214).

Given the mediating function of authority in Raz's service conception of authority, it is natural to suppose that authority is justified only insofar as it does a better job than its subjects of deciding what the reasons that antecedently apply to the subjects require by way of action.

To see the motivation for NJT, it is helpful to compare the justification for taking someone's advice. Consider a case in which one person *P* will be hurt if her friend *Q* does not accept *P*'s advice. The desire to spare *P*'s feelings might, depending on the circumstances, be a reason for accepting *P*'s advice; if the matter were sufficiently inconsequential and the advice were harmless, *Q* might be justified in accepting *P*'s advice to avoid hurting her feelings. But, as Raz points out, that is not the normal reason for accepting advice: "The normal reason for accepting a piece of advice is that it is likely to be sound advice" (Raz 1985, 19). Likewise, it seems natural to accept and follow a practical authority's directives because they are likely to require subjects to do what, as a matter of objectively right reason, ought to be done.[34] If so, NJT states, at the very least, considerations that are clearly relevant with respect to justifying authority.

Raz describes NJT as a "*moral* thesis about the type of argument which could be used to establish the *legitimacy* of an authority" (Raz 1985, 18; emphasis added). Further, on Raz's view, the legitimacy of an authority entails what he calls "a [moral] right to rule,"[35] and thus Raz believes that it entails the existence of a content-independent moral obligation on the part of subjects to obey the authority's directives

---

[33]This strikes me as a difficult issue. While a physician is clearly an epistemic authority, it is not clear whether a physician is a practical authority. On the one hand, a physician's recommendations do not seem to be directives in the relevant sense; on the other, it does not seem to be irrational for a patient to accept a physician as being a practical authority. But deciding this issue is not important here insofar as it is commonly accepted that there can be practical authorities that lack the capacity to coerce compliance or punish disobedience.

[34]For a critical discussion of NJT, see Himma (2007).

[35]One reason to think that physicians lack practical authority is that it is implausible to think that a physician has a "moral right to rule." Whether that is true depends, I suppose, on the character of a patient's acceptance. Nonetheless, it is worth noting that if NJT is construed as a sufficient condition for the justification for accepting someone as a practical authority, it would justify accepting

(provided, of course, that the content of the directives falls within the scope of the authority's legitimacy).[36]

There are a number of reasons to think that NJT fails as a general principle of moral legitimacy. To begin, even if NJT succeeds as a principle of moral legitimacy with respect to practical authority lacking authorized access to coercive mechanisms, it is reasonable to think that something more would be needed to justify the legitimacy of coercive state authority. The introduction of coercive enforcement mechanisms complicates the moral issue considerably because persons have a presumptive moral right to be free of coercion. The exercise of all practical authority raises a moral issue because it infringes a person's moral right (or duty, as Wolff would have it) to make and execute her own decisions. But coercion introduces the threat of violent repercussions for non-compliance that would appear to require much more by way of moral justification than a principle that merely requires that the authority be better than the subject at deciding what the subject should do according to right reason. Even if any person who functions as a practical authority is passively accepted or acquiesced to by the subject, it is not enough to justify the threat of violence that the authority knows better than the subject what the latter should do according to right reason.

But the potential problems do not end here. Satisfaction of NJT seems neither sufficient nor necessary to give rise to content-independent moral obligations to *obey* the directives of a de facto authority; otherwise put, satisfaction of NJT does not seem to be either sufficient or necessary for authority to be morally legitimate. Clearly, it is not sufficient. The mere fact that complying with the directives of an authority is more likely to conduce to my interests (assuming right reason dictates that I should do what conduces to my interests) than not complying can, to borrow from Hart, *oblige* me to do what the authority directs—"obliging" being a notion that is prudential in character; if complying would conduce to my interests, I would be foolish not to accept the authority's directives. But the mere fact that following an authority is in my interest cannot, in and of itself, morally *obligate* me to obey the directives of the authority.

The deeper problem here is a familiar one. Raz's NJT seems to take the satisfaction by P of whatever standards confer epistemic authority over Q as sufficient to confer practical authority on P over Q.[37] But if this is Raz's considered view, it is mistaken. Epistemic authority is concerned with providing truth-conducive reasons to believe

---

physicians as practical authorities. These conflicting considerations highlight the difficulties concerning the issue whether physicians are practical authorities. See note 33, above.

[36]The idea that a patient has a content-independent moral obligation to obey a physician seems highly implausible, which, of course, casts doubt on the idea that physicians are practical authorities. Although I think it useful to assume physicians are practical authorities to save space, nothing turns on this issue—if, as many theorists seem to think, there can be practical authorities that lack coercive authority.

[37]Heidi Hurd takes a somewhat different position—although its ancestral lines to Raz are clear; as she puts her position: "law can at best provide us with reliable moral advice, but cannot provide us with any reasons to do what morality otherwise prohibits" (Hurd 1999, xiii). She goes on to claim that law can have a form only of theoretical authority (a species of epistemic authority) and not practical authority. Whereas Raz seems to want to infer justified practical authority from justified

and does not obviously imply practical authority. *P*'s epistemic authority with respect to what *Q* should do according to right reason gives *Q* a reason to believe what *P* has said *Q* should do, but it does not necessarily give *Q* an exclusionary reason to do what *P* has said—and it certainly cannot, by itself, give rise in *Q* to a justified expectation that *P* complies. The mere fact that a physician, who has epistemic authority with respect to my physical health, informs me that (1) I should take an antibiotic because (2) I have a bacterial infection gives me a reason to believe both (1) and (2). But while that fact, *by itself*, might give me a reason to take the antibiotic, it is implausible to believe that the character of the reason is an exclusionary reason. Further, the idea that the physician's epistemic authority, alone, creates in the physician a morally justified expectation that I comply seems straightforwardly false. If a physician has practical authority over a patient, it cannot be explained in terms of the physician's epistemic authority alone. Something else in the physician–patient relationship will also be needed to explain the physician's practical authority.

Stephen Darwall makes a similar argument, although his reasoning differs in that it relies on his problematic conception of second-personal reasons. As he puts the point:

> Meeting the standards of the normal justification thesis is not, however, sufficient to establish practical authority. There are cases where one person might very well do better to follow someone else's directives where it seems clear that the latter has no claim whatsoever on the former's will and actions and consequently no practical authority with respect to him. And cases where an "alleged subject" would do better in complying with independent reasons where genuine authority does seem to be involved also seem to involve some assumed background accountability relation that gives the authority's directives standing as second-personal reasons. In these cases, it is the latter that establishes the directives' authority, not the former (Darwall 2009).

Darwall argues, in effect, that more is needed than just epistemic authority to give rise to practical authority over another person; in particular, what is needed is that the subject must be, for other reasons, antecedently accountable to the person who is the epistemic authority. Lacking those other accountability relations, epistemic authority is insufficient to give rise to practical authority. That much is surely correct.[38]

Darwall is surely correct in thinking that NJT fails as a sufficient condition for justifying practical authority. But his remarks above are nonetheless problematic inasmuch as they rely on a deeper account of practical authority that more accurately explicate the notion of *standing* than the notion of practical authority. The problem is that Darwall believes that the norms of morality, insofar as they protect us from certain acts, make each of us a practical authority over every person who is ultimately accountable to us. He believes that it is our status as practical authorities that explains

---

epistemic (or theoretical) authority, Hurd argues that the best law can provide is theoretical authority. It cannot have the kind of practical authority that would provide a content-independent exclusionary reason to do something that violates moral requirements.

[38] Scott Hershovitz argues that satisfaction of NJT is not a sufficient condition for legitimacy because matters having to do with whether a state employs democratic procedures are also relevant. See Hershovitz (2003), and note 14, above.

why we have a justified demand that obligations owed to us be satisfied and can make justified claims to that effect.

But this stretches the notion of practical authority well beyond its intuitive boundaries. We are not accurately characterized as practical authorities over others with respect to obligations owed to us insofar as we are not the source of those obligations—at least not in the sense of "practical authority" that conceptual jurisprudence wishes to explicate. If God, for example, manufactures morality, then God is a practical authority and God's directives are authoritative; that we enjoy the protections of these directives does not, on any remotely plausible conception of authority, entail that we are all practical authorities with respect to claiming satisfaction of those obligations. But if morality consists of objective truths that are independent of God's willings or commands, then the norms of morality are authoritative without there being a practical authority, if practical authority is, by nature, personal in character. As we have seen in the case of a legal system, there can be authoritative directives that do not originate with a practical authority; however, if some person's directives are authoritative, it must be because that person is the source of the directive and is a practical authority. While Darwall is correct in thinking that *P*'s epistemic authority over *Q* cannot give rise to practical authority over *Q* in the absence of preexisting norms making *Q* accountable, his views about what gives rise to these accountability relations—which stem from his problematic account of second-personal reasons and practical authority—are unhelpful in securing the point.

NJT is even less successful, given its content, in providing a moral justification for using coercive means to enforce those directives against me. No matter how much better a de facto authority might be than I am in discerning the requirements of right reason, that fact alone is not sufficient to justify using coercive force against me to ensure that I obey her judgments.

Nor does satisfaction of NJT seem necessary for an authority to be morally legitimate. Although consent-based theories of legitimacy generally hold that subject consent to authority is, by itself, sufficient for the legitimacy of an authority over that subject and hence that NJT is not necessary for legitimacy, this is too strong. My consent to follow authority, in and of itself, is not enough to give rise to a content-independent moral obligation to follow that authority. Insofar as the legitimacy of an authority over a person is wholly grounded in that person's consent to it, its continuing legitimacy depends on that person's continuing consent. The problem is that, on any ordinary conception of unilateral consent, a person's consent, since voluntary, can be withdrawn at any given moment—and it is as implausible to suppose that a person can give irrevocable effective consent to be bound by a state *for her entire life* as it is to suppose a person can give effective consent to being a slave.

Accordingly, if my moral obligation to follow the directives of an authority is grounded *entirely* in my unilateral consent, that obligation can be extinguished at any time by my withdrawal of consent. Indeed, on such an account, it would be difficult to distinguish the obligation that authority gives rise to form non-obligatory behaviors that I am free to engage in or refrain from at my disposal because I am always free to withdraw and renew my consent to authority at my discretion.

Consider, for example, the authority of an arbitration (which Raz takes as paradigmatic of practical authority): My unilateral consent, without more, to follow the directive of an arbitrator regardless of its content cannot morally obligate me to follow it. For if the authority of the arbitrator over me is *wholly* grounded in *my* consent, then the arbitrator's authority over me is terminated simply by my withdrawal of that consent.

What is missing is the reliance of others on my consent—and such reliance typically takes the form of a corresponding commitment to obey the directives of an authority. If you and I both consent to abide by an authority's directives, then we are both forgoing options that would otherwise be available. There are different ways to explain how this gives rise to a moral obligation on the part of each of us to obey the directives. One might take a strict contractarian view and conceptualize our joint consenting as constituting a contract that gives rise to the obligation. Or one might argue that it would be unfair to allow someone to reap a benefit from disobedience given that other people have abdicated such a benefit. But, however, this is done, a key element in the legitimacy of authority is typically thought to rest on the express or implied consent of all persons over whom the authority is legitimate.

Raz is, of course, aware of the importance of consent to most theories of legitimate authority, but he rejects the idea that it is consent that binds the individual: "[a]greement or consent to accept authority is binding, for the most part, only if conditions rather like those of the normal justification thesis obtain" (Raz 1994, 214). The key element of his normative theory of legitimacy is the superior expertise of the authority in determining what subjects ought to do according to right reason; subject consent is of little to no importance on Raz's view.

It is worth nothing that Raz gives little, if anything, by way of argument for this conclusion. Contractarian theories of moral legitimacy have been widely regarded as plausible for centuries. These theories go back as far as Hobbes and Locke, and they remain of considerable contemporary influence in theories that are as different as the theories of Rawls and Nozick. Given the tremendous influence that such theories have enjoyed, Raz needs more argument than he gives to reject the central importance of consent to legitimacy.

Certainly, there are limits on the extent to which consent gives rise to moral obligations. As Raz points out, there are certain restrictions on how an authority may decide what a subject must do—even if the subject consents. Consider the context of a legal system. A judge, for example, could not legitimately decide a legal dispute on the basis of a coin flip, again, even if the parties consent to the judges doing so. Likewise, mutual consent and reliance are not enough to rescue a bargain if it is sufficiently unfair—either because of the bargain's content or because it was not negotiated at arm's length.

But these are exceptional circumstances and not the general rule with respect to the relation between consent and authority. If the parties are capable of giving effective consent to authority and the consent is secured in a fair manner, then the conditions articulated by NJT are not necessary for consent to authority to create a moral obligation to obey the directives of that authority. If, e.g., one party foreseeably relies to her detriment on the other's consent to abide by the directives of an authority,

then the latter is bound by her consent, even if she gave her consent for reasons other than those described in NJT—indeed, even if she gave her consent for imprudent or ill-advised reasons. If this is correct, then NJT is neither a sufficient condition nor a necessary condition for the legitimacy of authority.

Raz would respond that he intends NJT as providing neither necessary nor sufficient conditions for justifying practical authority; his claim is the considerably weaker one that satisfaction of NJT is the "normal" way to justify that "[one] person should be acknowledged as having practical authority over another" (Raz 1994, 214). Accordingly, his response is that such criticisms misrepresent his position with respect to NJT.

There are two problems with this response. To begin, the claim that this is the *normal* way to justify practical authority is an empirical claim that would have to be supported with sociological evidence that has not been provided. Even so, there is little reason to think that people commonly cite epistemic authority as sufficient to justify practical authority, which is what the word "normal" connotes. It is surely true, as an empirical matter, that people commonly justify *A*'s accepting a piece of advice on the ground that its source, *B*, is more likely to be correct about what *A* should do. But practical authority differs from advice in morally salient ways. Unlike advice, for instance, practical authority involves a content-independent moral obligation to do what the authority's directives require, as well as a justified expectation on the part of the practical authority that the subject complies with the authority's directives. If the authority may use coercive enforcement mechanisms to enforce her directives, the problem becomes still more complex, from a moral standpoint. While Raz's model of advice is a useful heuristic device for understanding the motivation for NJT, the morally significant differences between being an advisor and being a practical authority are such as to require much more by way of justifying practical authority than by way of justifying being an advisor. Raz's theory of justifying authority seems problematic, no matter how it is construed—whether as providing necessary and sufficient conditions or as expressing the *normal* way to justify authority.

References

Christiano, T. 2013. Authority. In *The Stanford Encyclopedia of philosophy*, ed. E.N. Zalta. http://plato.stanford.edu/archives/spr2013/entries/authority/.

Darwall, S. 2009. Authority and Second-Personal Reasons for Acting. In *Reasons for Action*, ed. J. Wall, and D. Sobel. Cambridge: Cambridge University Press.

Dworkin, R. 2002. Thirty Years On. *Harvard Law Review* 115: 1655–1687.

Frankena, W. 1966. The Concept of Morality. *Journal of Philosophy* 63: 688–696.

Gert, B. 2012. The definition of morality. In *The Stanford Encyclopedia of philosophy*, ed. E.N. Zalta. http://plato.stanford.edu/archives/fall2012/entries/morality-definition/.

Gilligan, C. 1982. *In a Different Voice*. Cambridge, Mass.: Harvard University Press.

Green, L. 1989. *The authority of the state*. Oxford: Oxford University Press.

Hart, H.L.A. 1982. Commands and authoritative reasons. In *Essays on Bentham*, ed. H.L.A. Hart. Oxford: Oxford University Press.

Hershovitz, S. 2003. Legitimacy, democracy, and Razian Authority. *Legal Theory* 9: 201–220.

Himma, K.E. 2001. Law's Claim to Authority. In *Hart's postscript: essays on the postscript to the concept of law*, ed. Jules L. Coleman. Oxford: Oxford University Press.
Himma, K.E. 2007. Just 'cause you're smarter than me doesn't give you a right to tell me what to do: legitimate authority and the normal justification thesis. *Oxford Journal of Legal Studies* 27: 121–150.
Himma, K.E. 2016. *Can there really be law in a society of angels?* Available at SSRN: https://ssrn.com/abstract=2839942 or http://dx.doi.org/10.2139/ssrn.2839942.
Himma, K.E. 2017. The authorization of coercive enforcement mechanisms as a conceptually necessary feature of law. *Jurisprudence* 7: 593–626.
Hurd, H. 1991. Challenging authority. *Yale Law Journal* 100: 1611–1677.
Hurd, H. 1999. *Mortal Combat*. Cambridge: Cambridge University Press.
Kohlberg, L. 1984. *The Psychology of moral development: The nature and validity of moral stages*, vols. 1 and 2. New York, N.Y.: Harper and Row.
May, T. 1998. *Autonomy, authority and moral responsibility*. Dordrecht: Kluwer Academic Publishers.
Perry, S. 1989. Second order reasons, uncertainty, and legal theory. *Southern California Law Review* 62: 913–994.
Rosen, G. 2014. Abstract Objects. In *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. http://plato.stanford.edu/archives/fall2014/entries/abstract-objects/.
Raz, J. 1975. *Practical reason and norms*. Oxford: Oxford University Press.
Raz, J. 1985. Authority and justification. *Philosophy & Public Affairs* 14: 3–29.
Raz, J. 1994. *Ethics in the public domain*. Oxford: Oxford University Press.
Shapiro, S. 2011. *Legality*. Cambridge, Mass: Harvard University Press.
Simmonds, J.A. 2001. *Justification and legitimacy: Essays on rights and obligations*. Cambridge: Cambridge University Press.
Wolff, R.P. 1970. *Defense of Anarchism*. New York, N.Y.: Harper and Row.
Zagzebski, L.T. 2012. *Epistemic authority: A theory of trust, authority, and autonomy in belief*. Oxford: Oxford University Press.

The Authority of Law

Veronica Rodriguez-Blanco

Law transforms our lives in the most important way: it changes how we act and because of this it gives rise to fundamental questions. One such question concerns legal authority and individual autonomy and asks; if we are autonomous agents how do legislators, judges and officials have legitimate authority to change our actions and indirectly change how we conduct our lives? We conceive ourselves as active agents who determine how and when to act, and we conceive ourselves as the planners of our own lives and the creators of change. Law asks us, however, to perform actions that range from the trivial to the complex. Law requires us, for example, to stop at traffic lights; park our vehicles in specially allocated areas; exercise our professional judgment in a responsible and non-negligent manner; pay our taxes; recycle our rubbish. Law asks us to perform innumerable tasks, almost all of which we perform intentionally and in full awareness. But how is it possible for me to do, in full awareness, as the law asks and, at the same time, be in control of my own destiny? How is my free will affected by the law?

But how is this possible when I am simply trying to conform with what the law says? This means, I am trying to follow what the law says without giving much thought or without engaging my will or intention.

Legal and political philosophers have tended to examine legal authority and autonomy and have consequently put forward the following questions: (a) Can there ever be legitimate authority?; (b) What are the conditions of legitimate authority? and (c) Does the possibility of legitimate authority diminish or assuage the antagonism between authority and autonomy?

I find that posing the problem and the questions in this way is unsatisfactory because it presupposes what we need to explain; i.e. the nature of authority and

The entry relies on material previously published in Rodriguez-Blanco (2014b, 2017).

V. Rodriguez-Blanco (✉)
School of Law, University of Surrey, Guilford, UK
e-mail: v.rodriguez-blanco@surrey.ac.uk

G. Bongiovanni et al. (eds.), *Handbook of Legal Reasoning and Argumentation*,
https://doi.org/10.1007/978-90-481-9452-0_9

whether there is a "genuine" antagonism between autonomy and legal authority. Within this framework authority is given, and the starting point of the theorist is the following statement: If there is a legitimate authority then conditions *x*, *y* and *z* need to be fulfilled, but it is not shown how there is or whether there could be something such as legitimate authority. The received view begins by recognizing the phenomenological fact that legal officials and authorities issue commands and directives. It is usually said that if authorities have the right to command and addressees the duty to obey, then the officials have legitimate authority.

Theorists usually argue in favour of a particular political theory, for example liberalism or perfectionism, and engage with a set of key values, for instance expert knowledge or democratic values that provide the grounds for "rights" and "duties" and that enable us to grasp the conditions of legitimate authority. The traditional strategy, therefore, begins top-down from a plausible view on political theory that leads to the framework that justifies authority. There is no doubt that the traditional strategy has provided us with a rich understanding that has advanced our grasp of the normative conditions that make possible legitimate legal authority. However, the traditional strategy fails to provide a microscopic view of the phenomenon of legal authority and falls short of explaining how legal authority truly operates on individual human beings.

By contrast, the strategy of this study is to focus on the agent; i.e. the addressee of the legal command or directive who performs the action requested by the legal official. This strategy is bottom-up, from the level of agency and practical reason to the justificatory framework of authority. It also begins with the naive phenomenological observation that *X* commands *Y* to perform the action *p* (an action *p*-ing to *Y*). Thus, it is intelligible to us that *Y* performs the action *p* as requested by *X*. The key question that this study aims to investigate is how a legal command or directive, just because it is a legal command or directive, effectively changes the agent's course of action. A set of sub-questions arise: Does the command intervene in the practical reasoning of the agent or addressee? If this is the case, how does this intervention operate? Moreover, what are the limits of our phenomenological observations, in other words can I truly observe that you are performing an action because you are complying with a legal directive or command? What happens in the agent that enables her to comply with the legal command or directive? When we perform an action because we are complying with the legal command or directive, are we still active, self-governed autonomous agents? In what sense are we still autonomous agents? The task of this study is to explain what legal authority is and the premise of the study is that this question can only be answered through understanding of how legal authority operates upon the agent: if we recognize that legal commands or directives intervene upon, affect and change the agent's practical reasoning, then we need to understand and explain how this happens.

1 Human Action and Authority: Tracing the Correct Relationship

It is recognized by theorists and laymen that law is a social practice. However, if social practices are constituted by human actions then the following question arises: "What is the sound characterization of human action that enables us to provide a satisfactory explanation of the production of authoritative legal rules and directives?" The key feature of legal rules and directives is that they guide the behaviour of citizens and has a normative force on the addressees of legal rules and directives. It is, however, puzzling how human beings are able through their actions to produce such a complex state of affairs; i.e. a legal rule that is authoritative and intervenes in the reasoning and actions of the addressees of the rule.

Let us suppose that we explain human action as merely an empirical phenomenon, i.e. a set of regular patterns produced by the reason-beliefs or acceptance-beliefs of the participants, which are construed as mental states.[1] Within this framework of explanation, the authoritative character of the legal rules, their guiding role and normative force are utterly mysterious. For example, let us think about the legal rule that demands that citizens stop at red traffic lights and also about citizen "c" who does this numerous times every morning when driving to work. Following the empirical model of human action, the empiricist will say that citizen *c*'s action is explained by the fact that "there is a rule that is grounded on reasons that respond to what everyone does" (Lewis 1969); or, rather, "there is a rule that is grounded on our accepted reason-beliefs towards such a rule or accepted reason-beliefs towards a second-order rule about such a rule" (Hart 2012); or, even more, "there is a rule that is the result of deep conventions, which are the result of social practices, responsive to our social and psychological needs, arbitrary, grounded on a reason-belief to follow them, instantiated in superficial conventions and resistant to codification" (Marmor 2007. For a criticism of this view, see Rodriguez-Blanco 2016).

We feel, however, that there is something fundamentally missing in this purely empirical portrait of human action. It seems to imply that if one day citizen "c" decides not to do what everyone does or accepts, and decides instead not to stop at the red traffic lights, and consequently her vehicle collides with a number of other vehicles and she kills a child, then (following the empiricist explanation of human action) the only mistake she made in her reasoning that led her to the catastrophic action is that she did not accept what everyone accepted, or rather she did not have the appropriate reason-belief as mental state to follow the rule. This is a strange understanding of her reasoning, though it follows logically from an explanation of human action in terms of purely empirical features; i.e. social facts, beliefs or intentions as mental states, and reasons explained in terms of beliefs and therefore mental states.

[1] According to the empirical account of intentional action, the acceptance of legal rules provides reasons for actions in the context of the law. For a full explanation of the empirical account of action in the context of the law and its criticism, see Rodriguez-Blanco (2014b, Chap. 5). I argue that the empirical account of intentional action is parasitic on the "guise of the good" explanation of intentional action.

The explanation of the reasoning of the agent in empirical terms is equally unintelligible in examples where what is at stake is the life, dignity or another fundamental value that we human beings care about. Let us scrutinize the following example. If an official aims to enforce the court decision that has established that citizen "p" has violated the physical integrity of another citizen and therefore should be punished with imprisonment and we ask for an explanation of the official's coercive action, it would be puzzling to hear the following response: "Citizen "p" has violated a constitutional rule which is grounded on our acceptance-belief or reason-belief which lies behind the constitutional rule." This value-free or value-neutral response cannot truly explain why citizen "p" has to go to prison according to a court decision. Does it mean that if citizen "p" escapes from the coercion of the official and manages to leave the country, then the only mistake in her reasoning that leads her to flee the country is her disagreement with either the acceptance-belief that there is a valid constitution or secondary rule, or her disagreement with the acceptance-belief towards the constitutional rule and penal code that protects the physical integrity of all citizens? Thus, it is not that she disagrees with the value that is the content of the acceptance-belief or reason-belief, rather she disagrees with the acceptance-belief or reason-belief. The disagreement is just about beliefs, and therefore, according to the empirical account, the parties in disagreement are in different mental states. This is an equally strange and puzzling diagnosis of our disagreements.

When we characterize what legislators, judges, officials and citizens do in terms of actions as empirical phenomena, we seem to miss something fundamental. Worse, the empirical account of action cannot satisfactorily explain the guiding role of the law.[2]

Let us go back to our first example. Citizen "c" is a law-abiding citizen who aims to follow and be guided by the law, and on her journey to work, she knows there is a legal rule that states she ought to stop at red traffic lights. According to the empirical characterization of human action, she stops at red traffic lights because she has the acceptance-belief or reason-belief that there is such a rule and this acceptance-belief or reason-belief causes her to press the brake pedal on each relevant occasion. If she were asked why she presses the brake pedal she will reply, "because there is a red traffic light," and if she were asked, "why do you stop at the red traffic light?" she would reply, "because there is a secondary rule that is accepted by the majority of the population and this establishes the validity of the rule 'citizens ought to stop at red traffic lights'." Alternatively, she might reply, "I stop at the red traffic light because of the rule," but now the mystery is "why do you act according to the rule?," to which she might answer, "because rules give me reasons for actions." The empirical account explains reasons in terms of beliefs/desires as mental states (Davidson 1980), and

[2] Arguably, Raz's explanation of how legal rules intervene in our reasoning is non-empirical since he has emphasized that a reason for action should not simply be understood as beliefs as mental states (Raz 1979, 1986, 1999). However, in Rodriguez-Blanco (2014b, Chap. 8), I argue that Raz's explanation of legal authority is a theoretical explanation of our reasoning capacities; i.e. when we explain how legal directives and rules intervene in the citizen's practical reasoning from the third-person perspective. His explanation ignores the first person or deliberative point of view of the citizen who follows legal rules.

then it seems that it is the mental state that is causing the action. This is a problematic picture because it supposes that for each action I need to "remember" my belief/desire so that I am able to be in the right mental state so that I can stop at the red traffic light. However, we stop at red traffic lights even when we are tired or when we do not "remember" (Wittgenstein 1953, Section 645) that we ought to stop at red traffic lights, and therefore we somehow just "know how to go around" and stop at red traffic lights. Furthermore, the predominant empirical picture of human intentional action cannot explain the diachronic structure of intentional action. That is, we stop at red traffic lights over a prolonged period of time and even though the relevant mental state might be absent, we still continue doing it and it seems that we do it for a "reason" that tracks values or good-making characteristics.

Imagine that there is an emergency. Citizen "c" needs to bring her neighbour to the hospital because he is dying and consequently she decides not to stop at a red traffic light. Does this mean, if we follow the empirical account of human action, that in order to explain her action we need to say that she surely needed to "forget" that she had the relevant belief as mental state of "stopping at red traffic lights," or perhaps she decided "to get rid" of her acceptance-belief concerning the rule "we ought to stop at red traffic lights"? Or, perhaps, she "decided to suspend" her beliefs about the rule "we ought to stop at red traffic lights." In the two latter possibilities, we cannot say that it is not only a "belief" that plays a role in action, but rather the "will" of the rule-follower. She has used the words "get rid of" and "decided to suspend." It seems that there is something else going on. Imagine that we ask her, "why did you not stop at the red traffic light?" The empirical answer, "because I do not have the acceptance-belief or reason-belief towards the rule now for this specific instance," would be an odd one. Citizen "c" is more likely to say, "Don't you see it? My neighbour is dying and I want to save his life." Furthermore, if reasons for actions are belief-acceptance or if they give me a reason for action and this reason is merely a mental state, how can I be guided by rules and principles? If the empirical explanation of action is the sound characterization, then the guidance of rules and principles is effective because I am in the correct mental state. The entire work is done by my mental states as long as I am in the supposedly correct mental state. The deliberation of the legislator or judge and/or my own deliberation plays no role in the execution of my action of rule-following or principle-following. The content of the legal rule is irrelevant as long as the majority of the citizens are in the allegedly correct mental state.

Arguably, we need to resort to values in the form of good-making characteristics that are relevant to the specific form of life that is ours and that reflect what we care about individually and collectively. We need to understand human action in its naïve or fundamental form and this understanding, I argue, sheds light on the kind of things we produce, including human institutions such as law. Thus, if someone asks citizen "c" why she stops at the red traffic light on her way to work, there is a naïve explanation of her action that seems to be more primary than any other explanation. Thus, she might respond, "because I do not wish to collide with other vehicles and kill pedestrians." If we ask her, "why do you not wish to collide with other vehicles and kill pedestrians?," she most is likely to reply, "because I value my property, other

people's property, and life" and if we keep asking, "why do you value property and life?," she will respond, "because property and life are goods."

We have learned that a mistaken conception of human action can take us down misleading routes in our understanding of the nature of law and, more specifically, its pervasive, authoritative, normative and guiding role in our lives. The sound explanation of human action will illuminate how human beings produce law and will also shed light on the authoritative and normative features of law. In Sect. 2, I explain and defend a conception of human action that diverges from the standard empirical conception. In Sect. 3, I scrutinize the consequences of this conception of human action for our understanding of the nature of law and its authoritative and normative character.

2 Intentional Action Under the Guise of the Good

We will now concentrate on intentional human action as the paradigmatic[3] example of human action to shed light on the making of law by legislators and judges and the character of legal rule-following.[4]

In her book *Intention* (1957) Elizabeth Anscombe engages with the task of explaining intentional action along the lines of the philosophical tradition of Aristotle and Aquinas and identifies a number of key features that characterize intentional action. These features include:

(a) The former stages of an intentional action are "swallowed up" by later stages

Intentional action is composed of a number of stages or series of actions. For example, if I intend to make a cup of tea, I first put on the kettle in order to boil water; I boil water in order to pour it into a cup of tea. Because my action of making tea is intentional, I impose an order on the chaos of the world and this order is the order of reasons. Thus, I put on the kettle in order to boil water and I boil water in order to pour it into a cup. This is how I understand the sequence of happenings in the world that I, as an agent, produce or make happen. But, arguably, there could be an infinite number of series of actions; there could be a continuous infinite, or ceaseless, seamless web of actions. The question "Why?" can always be prompted: "Why are you making tea?" and the agent might reply, "Because it gives me comfort in the morning." There is, however, an end to the "Why?" series of questions and the end comes when the agent provides a characterization of the end or telos as a good-making characteristic. The action becomes intelligible and there is no need to ask "Why?" again. The end as the last stage of the "Why?" series of questions swallows up the former stages of the action and makes a complete unity of the action. Intentional actions are not fine-grained, they are not divisible into parts. Thus, parts

[3]For a defence of a conception of paradigms as the best methodology to understand social and human concepts see Rodriguez-Blanco (2003).

[4]I am using the term "rule-following" but the same explanation applies to principle-following. See Rodriguez-Blanco (2012, 2014a).

of series of actions are only intelligible because they belong to an order that finds unity in the whole.

(b) Intentional action is something actually done, brought about according to the order conceived or imagined by the agent

Intentional action is not an action that is done in a certain way, mood or style (Anscombe 1957, Section 20). Thus, it is not an action plus "something else"; i.e. a will or desire that is directed towards an action. Intention is not an additional element; e.g. an interior thought or state of mind, it is rather something that is done or brought about according to the order of reasons that has been conceived by the agent. Consequently, if the question "Why?" has application to the action in question, we can assert that the action is intentional. The prompting of the question "Why?" is the mechanism that enables us to identify whether there is an intentional action. Intentional action is neither the mere movements of our body nor the simple result of transformations of the basic materials upon which agency is exercised, e.g. the tea leaves, kettle, boiling water. It is a doing or bringing about that is manifested by the expression of a future state of affairs and the fact that the agent is actually doing something or bringing it about according to the order of reasons as conceived or imagined by the agent (Anscombe 1957, Section 21–22).

(c) Intentional action involves knowledge that is non-observational, but it might be aided by observation

What is the distinction between practical and theoretical knowledge? Let us take a modified version of the example provided by Anscombe (1957, Section 32). A man is asked by his wife to go to the supermarket with a list of products to buy. A detective is following him and makes note of his actions. The man reads in the list "butter," but chooses margarine. The detective writes in his report that the man has bought margarine. The detective gives an account of the man's actions in terms of the evidence he himself has. By contrast, the man gives an account of his actions in terms of the reasons for actions that he himself has. However, the man knows his intentions or reasons for actions not on the basis of evidence that he has of himself. His reasons for actions or intentions are self-intimating or self-verifying. He acts from the deliberative or first-person perspective. There is an action according to reasons or an intention in doing something if there is an answer to the question why. It is in terms of his own description of his action that we can grasp the reasons for the man's actions. In reply to the question "why did you buy margarine instead of butter?," the man might answer that he did so because it is better for his health. This answer, following Aristotle's theory of action[5] and its contemporary interpretations advanced by Anscombe provides a reason for action as a desirability or good-making characteristic. According to Anscombe, the answer is intelligible to us and inquiries as to why the action has been committed stops. However, in the case of the detective when we ask why did you write in the report that the man bought margarine, the answer is that it is the truth about the man's actions. In the case of the detective, the knowledge is theoretical, the detective reports the man's actions in terms of the

[5] Aristotle (1934). Nicomachean Ethics I. i. 2; III. V. 18–21. See also Aquinas, Summa Theologiæ. I–II, q8, a1, Kenny (1979), Pasnau (2002), Finnis (1998, 62–71 and 79–90).

evidence he has of it. In the case of the man, the knowledge is practical. The reasons for action are self-verifying for the agent. He or she does not need to have evidence of his own reasons for actions. This self-intimating or self-verifying understanding of our own actions from the deliberative or practical viewpoint is part of the general condition of access to our own mental states that is called the "transparency condition."[6] It can be formulated as follows:

(TC for reasons for actions) "I can report on my own reasons for actions, not by considering my own mental states or theoretical evidence about them, but by considering the reasons themselves which I am immediately aware of."

The direction of fit in theoretical and practical knowledge is also different. In the former case, my assertions need to fit the world whereas in the latter, the world needs to fit my assertions. The detective needs to give an account of what the world looks like, including human actions in the world. He relies on the observational evidence he has. The detective's description of the action is tested against the tribunal of empirical evidence. If he reports that the man bought butter instead of margarine, then his description is false. The man, by contrast, might say that he intended to buy butter and instead bought margarine. He changed his mind and asserts that margarine is healthier. There is no mistake here.

The idea that we accept from the internal point of view primary or secondary legal rules (Hart 2012) presupposes an inward-looking approach to action as opposed to an outward-looking approach. The latter examines intentional actions as a series of actions that are justified in terms of other actions and in view of the purpose or end of the intentional action as a good-making characteristic, e.g. to put the kettle on in order to boil the water, in order to make tea because it is pleasant to drink tea. The former examines the mental states that rationalize the actions; however, at the ontological level, arguably, it is mental states that cause the actions. The mental states consist of the belief/pro-attitude towards the action. If the "acceptance thesis" is the correct interpretation of Hart's central idea concerning the internal point of view towards legal rules, then criticisms that are levelled against inward-looking approaches of intentional actions also apply to the "acceptance thesis" (Hart 2012). The main criticism that has been raised against the idea that the belief/pro-attitude pairing can explain intentional actions is the view that it cannot explain deviations from the causal chain[7] between mental states and actions. Let us suppose that you intend to kill your enemy by running over him with your vehicle this afternoon when you will meet him at his house. Some hours before you intend to kill your enemy, you drive to the supermarket, you see your enemy walking on the pavement and you suffer a nervous spasm that causes you to suddenly turn the wheel and run over your enemy. In this example, according to the belief/pro-attitude view, there is an intentional action if you desire to kill your enemy and you believe that the action of killing your enemy, under a certain description, has that property. Ontologically, the theory would establish that you had both the desire to kill your enemy and the belief

[6] See Evans (1982), 225 and Edgeley (1969). The most extensive and careful contemporary treatment of the "transparency condition" is in Moran (2001).

[7] The first person to discuss deviant causal chains was Chisholm (Chisholm 1976).

that this action has the property "killing your enemy." Thus, this mental state has caused the action and there is an intentional action. The problem with this view is that it needs to specify the "appropriate causal route." Davidson has made much effort to specify the "attitudes that cause the action if they are to rationalize the action" (Davidson 1980, 79). In the following paragraph, Davidson seems to fear that the idea of attitudes causing action might lead to infinite regress:

A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to lose his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally. It will not help, I think, to add that the belief and the want must combine to cause him to want to loosen his hold, for there will remain the two questions how the belief and the want caused the second want, and how wanting to loosen his hold caused him to loosen his hold.

Here we see Davidson struggling with his own proposal (for an illuminating discussion of this point see Vogler (2007)). He asks how attitudes must cause actions if they are to rationalize actions. Davidson's model of intentional action does not help us to determine whether there is an intentional action, it only helps us to determine the conditions that would explain the existence of an intentional action. The intentional action is already given. A similar criticism is applicable to the "acceptance thesis" and to this we now turn.

Let us suppose that I intend to go to the park in my car however I read a sign at the entrance of the park that states "Vehicles are not allowed to park in the park," I turn the wheel of my vehicle, reverse it and park a few streets away. You ask me why I turned the wheel of my vehicle, reversed and parked a few streets away from the park; I answer that I carried out these actions because there is a rule that states "Vehicles are not allowed to park in the park." According to the "acceptance thesis," my desire to follow the pattern of behaviour indicated by the rule and my belief that turning the wheel of my vehicle, reversing it and not parking in the park is the type of action or pattern of behaviour indicated by the rule. However, let us suppose that I desire to avoid parking in the park and have the respective belief. In other words, I accept "not parking in the park." On my way to the park, however, while following directions to the park, I take a wrong turning and end up parking just outside the park entrance. Even though the two criteria of the "acceptance thesis" have been met, this was not a case of following the legal rule by acceptance since I comply with the rule by accident.

The problem with the "acceptance thesis" is that it does not consider the action from the deliberative point of view; i.e. as it is seen from the point of view of the agent or deliberator. When the agent explains his actions he does not examine his own mental actions, rather he looks outwards to the vehicle, the park, the sign and so on. The reasons for actions, i.e. turning the wheel to reverse the vehicle, then parking outside the park to follow the rule, are self-evident or transparent to him. But then, an objector might advance, what is the good-making characteristic of a rule that, as in the example of the shopper who intends to buy margarine because is healthier, is the goal of the action of avoiding parking in the park. My reply is as follows. When

the driver is asked why he or she is turning the wheel and reversing the vehicle, his answer will be "because it is the rule." But this is still not completely intelligible unless we assume or know that the driver is a law-abiding citizen or that he believes in the general fairness of legal rules, etc. We can still ask him, "Why, because of the rule, do you do this?" His answer would need to be in terms of reasons as good-making characteristics for him, in order to make intelligible his intentional action. He will probably reply that he has reasons to follow the legal rule because it is the best way of preserving the peace of the park, or that he has reasons to follow legal rules in general because it is the best way of preserving coordination[8] among the members of a community. In a nutshell, the agent or deliberator needs to provide the reasons for the action in terms of good-making characteristics and the end or reason of the action provides the intelligible form of the action. This explanation of action has also been called a naïve explanation of action as opposed to a more sophisticated explanation of action; i.e. in terms of mental states.

If I am an agent that acts in an intentional way, I know that I am bringing about something and I know this without the need to observe every single step of my series of actions to verify that (effectively) I am acting (Anscombe (1957), Section 28–29). In performing my action I might be aided by observation, but I know what is the order of the series of actions and why. This is the essence of practical knowledge. You do not need a theoretical stance towards yourself, a verification and observation of the movements of your body to know that you are performing an intentional action and bringing about something. Following the previous example, you do not need to observe that "you are making tea" to know that you intend to "make tea" and that you are bringing this about. You put on the kettle and boil the water, you do not ask yourself, "let me see what my body is up to, let me observe what I am doing," and then infer from the movements of your body that you are actually bringing about "making tea." Of course you can be aided by observation, you need your sight to put the kettle in the right position and to pour the boiling water without spilling it. But you do not use your observation and inferences from the observational data to know that you are making tea.

The state of affairs that you intend to bring about is at a distance, it might not be within your sight (Anscombe 1957, Section 29–30). Imagine a painter who intends to make a painting. He has an idea about what the painting will look like, e.g. how the colours will be distributed across the canvas, and what topics and concepts will be at work in the painting. The painting is at a distance and the painter does not need to observe the movements of his body and the motion of the brushes to know what he is painting and why he is painting what he is painting. Certainly, his sight will help him to find the adequate colour at the correct time and to shape the figures at the right angle, but his intentional action is not what he observes; it is not the result of his painting but what he is actually doing. We do what happens.

(d) In acting intentionally, we exercise our practical knowledge. We can understand practical knowledge if we understand the structure of practical reasoning

[8] See Anscombe (1981) for an argument of authority as practical necessity.

Intentional action is not in the mind; it is not primarily a mental state; it is not an internal thought (Anscombe 1957, Sections 21–22, 25, 27–28). Rather, it manifests itself publicly and within the public reasons that we share as creatures with certain constitutions and belonging to a particular time and place. For example, we eat healthy food because it is good to survive; we look after our family because we love them; we avoid harm because we aim to enjoy pleasant things and so on. Similarly, we know that to make a cake you need flour, sugar, eggs and milk. If I see you mixing grass and earth and you tell me that you are making a cake, then I can assert, if I consider that you are in sound mind (your full capacities), that there might be a mistake in your performance or that you do not understand what it is "to make a cake."

According to Anscombe, Aristotle establishes a strong analogy between practical and theoretical syllogism and this has led to misinterpretations about what practical syllogism is (Anscombe 1957, Sections 33, 33–34). Like theoretical syllogism, practical syllogism is often systematized by Aristotelian interpreters as having two premises, i.e. major and minor, and a conclusion. It is said that, as in the case of theoretical syllogism, the practical syllogism is a proof or demonstration. The typical form might be as follows:

Vitamin X is good for all men over 60
Pig's tripe is full of vitamin X
I am a man over 60
Here is pig's tripe

But in this case nothing seems to follow about doing anything. Furthermore, the practical syllogism is sometimes interpreted as having an ethical or moral character and establishing a way to prove what we ought to do. Following the previous example, the conclusion might be, "I should eat pig's tripe." Anscombe rejects this view since Aristotle's examples are not in ethical contexts, i.e. "dried food is healthy," "tasting things that are sweet" that are pleasant. Additionally, the word "should" (*dei*) as it appears in the Aristotelian texts has an unlimited number of applications and does not necessarily refer to the ethical or moral context (Anscombe 1957, Section 35).

Aristotle insists that the starting point of any intentional action is the state of affairs or something that the agent wants and is wanted because it is presented to the agent as having good-making characteristics or as being valuable. For example, the man wants to have vitamin *X* because it is healthy. Furthermore, the practical syllogism is not limited to two premises and a conclusion; there can be many intermediate instances that are part of the syllogism. After a close analysis, the analogy between practical and theoretical syllogism breaks. Unlike theoretical syllogism, practical syllogism is not a proof or demonstration of a true proposition, nor is it a proof or demonstration of what ought to be done or what we ought to do. It is a form of how and why we are bringing something about when we are actually bringing it about.

Anscombe presents us with an alternative analysis to the practical syllogism and a different way to understand practical reasoning. Thus, the series of responses to the question "Why?" manifests or reveals the practical reasoning of the agent and enables us to identify whether the action that the agent is performing is intentional or

not. However, she warns us, the why-question methodology is as "artificial" as the Aristotelian methodology of practical syllogism (Anscombe 1957, Section 41–42). When we act intentionally, we are exercising a kind of reasoning which is not theoretical and which is grounded on a desire for that which seems to the agent to be constituted by good-making characteristics. You know the thing or state of affairs that you are bringing about because you desire the thing or state of affairs that you are bringing about, and you are able to desire the thing or state of affairs that you are bringing about because you know practically the state of affairs. Your desire arises because you represent the thing or the state of affairs to be brought about as valuable or good. Volition and knowledge do not fall apart (Anscombe 1957, Section 36). For example, if you are a painter, you know how and why the shapes and colours on the canvas are what they are, it is because you desire and value the painting you will produce that it should be such and such a colour and shape. But it is also true that because you desire and value this and not that arrangement of colours and shapes that you are able to know it practically. Consequently, moral approbation is irrelevant for practical reasoning and for our practical engagement with the world (Anscombe 1957, Section 37–38). This does not mean that there are no instances of objectively justified reasons for actions. On the contrary, we aim at getting it right and finding the genuine good-making characteristics that will provide meaning and intelligibility to the movement of our bodies. Therefore, the possibility of hitting the target of genuine good-making characteristics resides in our good characters and capacities. But to understand the basic structure of practical reason and the different scopes of agency, we do not need to begin from fully justified and objective values.[9]

Whatever strategy we follow to show the structure of intentional action, whether we take the Aristotelian practical syllogism or the Anscombian series of actions revealed by the question "Why?," we are able to grasp the mechanism of practical reasoning in its different manifestations.

In this section, I will argue that if Anscombe is right and both strategies are "artificial" ways of understanding (Anscombe 1957, Section 41–42), then a deeper and more "natural" way of understanding practical reasoning is by grasping the nature of the capacity that is exercised by the agent. In other words, the answers to the "Why?" questions show a capacity that the agent is exercising when acting. In the next section, I will show that the Aristotelian potentiality/actuality distinction sheds light on understanding the exercise and nature of our practical reasoning capacities. Furthermore, the potentiality/actuality distinction illuminates each of the key features of intentional action (*a*, *b*, *c,* and *d*) and their interplay as identified by Anscombe.

[9]In Chapter 9 of Rodriguez-Blanco (2014b), I show that robust value realism is indispensable to making sense of our actions, practices and first-order deliberative phenomenology. See Chap. 3 for a full defence of the "guise of the good model." See also Grisez's interpretation of Aquinas's precepts of natural law in Grisez (Grisez 1969, 368).

2.1 Aristotle's Distinction Between Actuality and Potentiality

Contra Parmenides, who argued that motion is impossible since something cannot come from nothing, Aristotle advances the idea that motion or change is possible if there is an underlying nature or constant feature that does not change. To explain this, Aristotle resorts to the distinction between potentiality and actuality. In *Metaphysics*, *book Θ*, Aristotle uses the analogical method to show that particular instances of the scheme or idea of potentiality and actuality[10] have a pattern.[11] Thus, he begins with the particular instances of capacity/change and matter/form to explain the common patterns that will illuminate the general scheme of potentiality/actuality. However, since our purpose is to elucidate the character of practical reasoning which is a power or capacity, and I have argued that the general scheme of potentiality/actuality will help us to clarify the nature of practical reason, it is circular to resort now to a particular instance of capacity/change to explain potentiality/actuality. I will, therefore amend the Aristotelian argumentative strategy and explain the general scheme of potentiality/actuality. I will then proceed to explain the particular instance of exercising our practical capacities as the actuality of a potentiality.

It is difficult to capture what "motion" is and many definitions of "motion" tend to use terms that presuppose motion (e.g. "a going-out from potency to act which is not sudden," but "going-out" presupposes motion and "sudden" is defined in terms of time which is also defined in terms of motion). Therefore, this kind of definition is discarded by Aristotle for being circular and unhelpful. Nor can we define motion in terms of pure potency, because if we say that "bronze is potentially a statue," we are merely referring to the piece of bronze which has not yet been changed and therefore there is no motion. You can neither refer to motion nor to change as what is actual. For instance, you cannot refer to what has been built or transformed, e.g. a building or statue, because it is not being moved, but has already moved. In the example of a building, the bricks, wood, clay, cement of the building have been already moved; and in the case of a statue, the bronze has already been transformed. Thus, Aristotle defines motion as a kind of actuality which is hard to grasp. In other words, the actuality of what exists potentially, in so far as it exists potentially (Aristotle, *Physics*, III.1.201a9–11). Motion is an actuality that is incomplete. It is hard to grasp and the tendency is to say that motion is the actuality. In the example of the house, it is the house that has been built. The other tendency is to say that motion is the privation of something; i.e. the going from nothing to something, from not being a house to being a house. Finally, the tendency is also to think that motion is what exists before potentiality, e.g. the bricks, steel, wood, cement. Contrary to these tendencies, Aristotle insists that motion is what happens exactly at the midpoint,

[10] I use this term as Kosman and Coope interpret it from Aristotle's *Physics, Books III and IV*. This means, the change that acts upon something else so that this something else becomes *F*; i.e. the fulfilment of a potentiality. For example, the building of a house by a builder so that the house becomes built. See Kosman (1969) and Coope (2009).

[11] I follow the interpretation of Aristotle's *Metaphysics book Θ* advanced by Frede (1994) and Makin (Aristotle 2006, 133). Cf. also Ross (1995).

neither before when nothing has been moved and is mere potentiality, and neither after, when something has been moved. Furthermore, motion is not privation; it is rather constitutive actuality. For example, if the baby has not learned to speak English, we say that the baby is potentially a speaker of English, when a man knows how to speak English and is in silence, he is also potentially a speaker of English, and finally when the man is speaking English, we say that he is actually an English speaker speaking English. However, the potentiality of the baby ($p1$) is different from the potentiality of the man in silence ($p2$), and motion is located in the second potentiality ($p2$), when the man is in silence, but begins to pronounce a sentence to speak English. Motion is midway and is not privative, but rather constitutive. We do not say that the man speaking English went from being a non-speaker of English to a speaker of English, we say that he spoke English from being in silence (he knew how to speak English, but did not exercise his capacities).

The previous example locates us in the domain of the particular instance of capacity and change as exemplified by the potentiality/actuality distinction. Aristotle argues that there are many different types of capacity, i.e. active/passive, non-rational/rational, innate/acquired, acquired by learning/acquired by practice and one-way/two-way capacities. Two-way capacities are connected to rational capacities, whereas one-way capacities are linked to non-rational capacities. For example, bees have a natural capacity to pollinate a foxglove flower in normal circumstances (Makin in Aristotle (2006), 43), ("normal" circumstances might include a healthy bee in an adequate foxglove, and the absence of preventive circumstances). In the case of two-way capacities there ought to be an element of choice or desire to act, and the rational being can exercise her capacity by producing or bringing about "p." Furthermore, she also knows how to produce or bring about "non-p." The paradigmatic example used by Aristotle is medical skill. The doctor knows how to make the patient healthy (p) and how to provoke disease or illness (non-p). Therefore, the doctor can bring about two opposite effects (Aristotle, *Metaphysics, Book Θ*, 1046b 4–5, 6–7). For Aristotle, to have a rational capacity is to have an intellectual understanding of the form that will be transmitted to the object of change or motion. Thus, the doctor will have an understanding of what it means to be healthy and without illness, but also of what it means to be ill. Let us suppose that a doctor is producing illness in the enemies through prescribed drugs. She needs to understand the order of the series of actions that will result in sickness for the enemies and she needs to possess knowledge about the necessary drugs to make the enemies to collapse. Her action will be directed to produce illness. But the doctor can choose otherwise, e.g. she can choose to make the enemy healthy.

In the exercise of practical reason, we choose to act (Aristotle, *Metaphysics, Book Θ*, 5, 1048 a10–11) and this choosing activates the action and directs the capacity towards the series of actions that will be performed. By contrast, a non-rational capacity is non-self-activating; its acts are necessary. If the bee is in good health and there are no obstacles, it will pollinate the foxglove flower. By contrast, rational agents need to choose or decide to act to produce a result.

When we say that the medical doctor has the rational capacity to change the unwell patient into a healthy human being, we say that she has the "origin of change." She is

curing the patient and therefore she is in motion because she actualizes her practical reasoning capacities to bring about the result as she understands it. She has an order of reasons that connects a series of actions and knowledge of how to produce changes.

She is the origin of change because her medical knowhow explains why certain changes occur in situations involving that object, e.g. the patient who suffers chickenpox has fewer spots and less fever. For example, when a teacher intends to teach and starts to say some sentences on the topic of "Jurisprudence" to her pupils, we say that she is teaching. She is the origin of change in the pupils who are the objects of change. Thus, the students begin to understand the topic and have a grasp of the basic concepts.[12] Similarly, when legislators create the law and judges decide cases, they establish rules, directives and principles and these rules, directives and principles can be found in statutes and case reports. Can we say that legislators and judges have reached the end of the process? No, we cannot: statutes and case reports do not represent the end of the process since citizens need to comply with the legal rules and directives and perform the actions as intended by the legislators and judges. We say that legislators and judges are the origin of change because they know how and have an order of reasons that enables citizens to comply with legal rules and directives. The order or reasons as good-making characteristics ground the rules, decisions and legal directives. In parallel to the situation of the teacher, I cannot say that I am teaching unless my pupils begin to understand the topic that I am teaching. Thus, the legislator cannot say that she is legislating and the judge cannot say that she is judging, in paradigmatic cases, unless there is some performance of their actions by the addressees as they intend.

The distinction between potentiality/actuality clarifies the structure of practical reason as a capacity that is actualized when we act intentionally. We can now understand that the features of an intentional action identified by Anscombe can be illuminated by the potentiality/actuality distinction. The idea that the former stages of an intentional action are swallowed up by the later stages is explained by the idea that motion is constitutive and not privative. It is not that when I begin to act I do so as an irrational or a rational being, and that I when finish acting I am a rational being, or that I go from non-intentional to intentional action, but rather that I go from being a rational being and potentially intentional action to being a rational being and actual intentional action. Later stages begin to actualize something that was potentially there. My practical reason was always there potentially and the intentional action actualizes an order of ideas provided by my practical reason. For Anscombe, intentional action is something actually done, brought about according to the order conceived or imagined by the agent. If practical capacity is understood in the light of the general scheme of actuality/potentiality, then intentional action involves knowledge that is non-observational, but it might be aided by observation. In acting intentionally, I am exercising my practical reasoning capacity and this capacity is in motion. This motion is represented at the midpoint; after I potentially have an

[12]Makin argues that the teacher analogy is intended to show that the teleological perspective is equally appropriate for other-directed capacities and self-directed capacity. See Aristotle (2006), 198.

intention to act and before I have reached the result of my intentional action. It is not that the forming of an intention from nothing to something is a magical process. It is rather that I potentially have the power to intend which in appropriate circumstances can be exercised. As being in motion, I am the agent who knows what she is doing and why she is doing what she is doing, but if I observe myself doing the action, then I have stopped the action. There is no action. There is no more motion and no exercise of my capacities. Finally, Anscombe asserts that in acting intentionally, we exercise our practical knowledge. Because we are the kind of creatures that we are, we can choose or decide to bring about a state of affairs in the world and we do this according to our order of reasons. Practical knowledge is potentially in all human beings and when we decide to bring about a situation or do certain things, then we actualize this potentiality. We can direct our actions to produce either of two opposing results, e.g. health or illness, ignorance or knowledge, as opposed to non-rational creatures who can only produce one result under normal circumstances and with no impeding conditions, e.g. the bee pollinating the foxglove. It should be noted that to have an actual capacity, such as practical reasoning and the capacity to act intentionally, does not mean that A can Φ, nor that A will Φ if there are normal conditions and no impending elements. Instead, it means that A will Φ unless she is stopped or prevented. Thus, once our practical reasoning capacity begins to be actualized, it will strive to produce or do what A (she) has conceived. Once A (she) decides or chooses to act, then a certain state of affairs will be produced unless she is prevented or stopped. Intentional action and practical reasoning are not dispositions like being fragile or elastic, nor are they possibilities that something will be done. They are powers.

Now that we have grasped the idea of potentiality/actuality as the general scheme for explaining the structure of practical reason, we can turn to the rule-compliance phenomenon and the creation of legal rules by legislators and judges, which raises a different set of difficulties that will be dealt with in the next section.

3 Law and *Energeia*: How Do Citizens Comply with Legal Rules?

So far we have argued that an intentional action is the bringing about of things or states of affairs in the world. We can argue, too, that there are different kinds of bringing about. Human beings can produce houses, clocks, tables, teacups and so on, but we can also produce rules of etiquette, rules for games, and legal directives, rules, and principles. Legislators create legal rules and directives and judges create decisions according to underlying principles and rules. These legal rules and directives are directed to citizens for them to comply with. They are meant to be used in specific ways. When a legislator creates a rule or a judge reaches a decision that involves rules and principles, she creates them exercising her practical capacities with the intention that the citizens comply with them. But how is this compliance possible? How do

legislators and judges create legal rules and directives that have the core purpose of directing others' intentional actions and of enabling them to engage in bringing about things and states of affairs in the world? In other words, how do other-directed capacities operate? This is the question that we aim to explore in this section.

Let us give two examples of authoritative commands to highlight the distinction between different kinds of authoritative rules:

Scenario 1 (REGISTRATION): you are asked by a legal authority to fill in a form that will register you on the electorate roll.

Scenario 2 (ASSISTANCE AT A CAR ACCIDENT): you are asked by an official to assist the paramedics in a car accident, e.g. to help by transporting the injured from the site of the accident to the ambulance, to assist by putting bandages on the victims, to keep the injured calm.

Arguably, the performance required by the addressee is more complex in the latter example than in the former since the latter requires the engagement of the will and the performance of a series of actions over a certain period of time, and it requires that the addressee should circumvent obstacles to achieve the result according to what has been ordered. It requires that the addressee exercises her rational capacity in choosing this way rather than that way of proceeding. While the addressee executes the order she needs to make judgments about how to do this or that. Successful performance as intended entails knowledge about how to proceed at each step in order to perform the series of actions that are constitutive of what has been commanded. This cannot be done unless our practical reasoning and intentional action are involved in the performance. In other words, the successful execution of the order requires the engagement of practical reasoning and therefore of our intentions. Furthermore, it requires an understanding of the telos or end as a good-making characteristic of what has been commanded. In the case of ASSISTANCE AT A CAR ACCIDENT, it requires engagement with the health and well-being of the victims of the accident. Thus, the addressee needs to know that the bandage ought to be applied in this way and not that way in order to stop the bleeding, and she knows that she needs to stop the bleeding in order for the victim to have the right volume of blood in his body. The victim needs a certain volume of blood in his body in order to be healthy and being "healthy" is something good and to be secured.

Because our practical reasoning capacity is a two-way capacity the agent needs to decide or choose to actualize this capacity which, prior to actuality, is mere potentiality. As in our example in Sect. 2.1, the speaker needs to decide or choose to speak in order to actualize their potentiality of speaking English. Then the exercise of their capacity to speak actualizes according to a certain underlying practical knowledge, e.g. the order of the sentences, grammar, style. It is not the case that as a bee pollinates a foxglove without any decision or choice by the bee, the agent will speak English and actualize their potential capacity to speak. In the case of legal rules, the question that emerges is how a legislator or judge can produce or bring about something that will engage the citizens' intentions so that they comply with legal rules or directives that are constituted by a complex series of actions. The core argument is that legislators and judges intend that citizens comply with legal directives and rules, and this intention is not merely a mental state that represents accepted reasons or reason-

beliefs. On the contrary, for the legislators' and judges' intentions (i.e. to engage the citizens' practical reasoning) to be successful, they need to exercise their own practical reason. It is not that they interpret or construct the citizens' mental states and interior thoughts so that their values and desires can constitute the ground that enables legislators, judges and officials to construct the best possible rules, directives or legal decisions according to the citizens' values as represented in their beliefs. On the contrary, they will look outward to what is of value and why certain states of affairs and doings are valuable (see the discussion on practical knowledge as non-observational Sect. 2, c). Reasons for actions as values and goods that are the grounds of legal rules and directives will engage others' practical reason. Therefore, the citizens' practical reasoning power or capacity become an actuality. If, as I have argued, our intentional actions become actuality by an order of reasons in actions and for actions that are ultimately grounded on good-making characteristics, then legislators and judges need to conceive the order of reasons as good-making characteristics that will ground their legal rules, legal directives and decisions. Judges and legislators would hence take the first-person deliberative stance as the privileged position of practical reasoning to disentangle what good is required and why it is required. In other words, if as judge or legislator you intend that your legal rule or directive is to be followed by the addressees and, arguendo, because these legal rules and directives are grounded on an order of reasons, then you cannot bring about this state of affairs, i.e. rule-compliance, without thinking and representing to yourself the underlying order of reasons. Let me give a simple example. You are writing an instruction manual on how to operate a coffee machine. You need to represent to yourself a series of actions and the underlying order of reasons to guide the manual's users. If you are a person of certain expertise, e.g. a manufacturer of coffee machines, then the practical knowledge that entails the underlying order of reasons is actualized without much learning and thinking. The required operating instructions are actualized as a native English speaker speaks English, after being in silence. By contrast, if you have only just learned to write instruction manuals for coffee machines, then you need to ask yourself "Why do it this way"? at each required action to make the machine to function. This process guarantees understanding of the know-how to operate the machine, and the success of the manual is measured by the fact that future buyers of the coffee machine are able to operate it. When legislators and judges create legal directives and legal rules they operate like the writers of instruction manuals, though at a more complex level. They need to ensure that the addressees will decide or choose to act intentionally to comply with the legal rules or directives and thereby bring about the intended state of affairs. But they also need to ensure that the order of reasons is the correct one so that the intended state of affairs will be brought about by the addressees. We have learned that the early stages of an intentional action are "swallowed up" by the later stages and ultimately by the reason as a good-making characteristic that unifies the series of actions. Thus, for addressees with certain rational capacities and in paradigmatic cases, understanding the grounding reasons as good-making characteristics of the legal rules and legal directives will enable them to decide or choose to comply with the rule and will guide them through the different series of actions that are required for compliance with the rules and directives.

Legal rules and directives do not exist like houses, chairs, tables or cups of tea. We need to follow them for them to exist. But we create legal rules and directives as we create houses, chairs, tables. We bring these things about by exercising our practical capacity and we are responsive to an order of reasons as good-making characteristics that we, as creators, formulate and understand. Thus, builders create houses that are either majestic or simple, elegant or practical, affordable or luxurious. To achieve the intended features of a house, builders need to select specific materials and designs, hire skilled workers, and so on. Similarly, legislators, officials and judges create legal directives and rules to pursue a variety of goods, e.g. to achieve safety, justice, the protection of rights. Legislators, officials and judges actualize their practical reasoning by creating an order of reasons in actions that will ground rules so that we are able to comply with them because we actualize our practical reasoning. Like builders, legislators, officials and judges need to choose values, goods and rights that will be fostered or protected by their rules or directives. Likewise, they need to formulate legal rules and directives that will have appropriate sanctions and are clearly phrased and followed procedures for their publicity. Arguably, what is at stake is not the mere publicity of a rule, but the publicity of the values that are embedded in the set of legal rules and principles. In this way, judges make the addressee of a directive choose or decide to actualize their potential practical reasoning capacity to comply with legal rules and directives. The addressees of a legal directive or rule are not like bees, who without decision and, given normal conditions and the absence of impediments, will pollinate the foxglove. As addressees of legal directive and legal rules, we need to choose or decide to bring about a state of affairs or things which are intended by the legislator, official or judge. This can be summarized as the idea that legal authority operates under the guise of an ethical-political account since it needs to present legal rules and directives as grounded on reasons for action as good-making characteristics.

As rational creatures, we are responsive to reasons as grounded in good-making characteristics, but if this is truly the case, how do mere expressions of doing as brute facts, such as "because I said so," or beliefs, intentions or reasons construed as mere mental states make possible the actuality of our practical reason? In fact, this is only possible if "because I said so" involves reasons in action that are grounded in good-making characteristics, e.g. "I am the authority and compliance with the authority has good-making characteristics." For example, compliance with authority is a secure way that some goods—apparent or genuine—will be achieved. The potentiality/actuality and capacity/change discussion show that as intellectual and rational beings, we need to apprehend the "form" that underlies the brute fact "because I said so," so that we are able to comply with legal directives and rules. As theoreticians, we now understand the limits of the empirical explanation of action, i.e. it has no "form" that makes intelligible the actuality of our practical reason and explains the dynamic reality of our intentional actions. Of course, we can decide that there is no such a thing as practical reason and that it is perfectly reducible to theoretical reason,[13] but

[13] See Enoch (2011) for a recent defence of the reductive approach. See Rodriguez-Blanco (2012) for a criticism of his position.

then the price we pay for this simple approach is too high: it leaves a set of human actions and the phenomenology of our first order or deliberative stance in the mists of mystery.

The "form" takes the shape of goods and values that are intended to be achieved by legislators, officials and judges. If it were a matter of mental or social facts, and we were able to apprehend the brute fact "because I said so" by our senses, or access legislators' and judges' reasons and values via our mental states only, without directly engaging with values and reasons, then how could we control and direct the doings and bringing about that are intended by legislators and judges? Some stages of the action will seem this and other stages will seem that. There is no way to bring about this and not that. Let us take the example of ASSISTANCE AT THE CAR ACCIDENT. I assist the official at the car accident because he has said so. I have no reason to assist him at the car accident; my action is only caused by my fear of sanction; i.e. a psychological impulse in me. But now as I am merely guided by my senses, it seems to me that I need to put the bandage on in this way rather than that way, but my sight alone cannot guide me on this. Since I am guided by my eyes and other senses, I do not know why I should apply the bandage or how I should apply the bandage. Furthermore, how can we attribute responsibility as we cannot be blamed for not "seeing" or "hearing" appropriately? By analogy, mere scribbles on the board by the teacher cannot make the pupil understand the topic that the teacher is teaching. The teacher needs to make transparent the premises and conclusions of her arguments so that the pupils can "grasp" the form of the argument and can themselves infer its conclusion.

Let us return to our initial example. Citizen "c" stops at the red traffic lights on her way to work. If we ask her "why are you stopping at the red traffic lights?" and we are satisfied with the empirical explanation which is, "because there is a secondary rule that is accepted by the majority of the population and this establishes the validity of the rule 'citizens ought to stop at red traffic light'," then how can we attribute responsibility to citizen "c," who just happen to have certain mental states? How can citizen "c" produce the required action just by remembering her mental state? By contrast, within the framework of the notion of practical reason that we have defended in this article, she will naïvely reply, "because the legal rules say so," and to reach intelligibility we could continue by asking, "why do you follow what the legal rules say?." She could then naïvely reply, "because I do not wish to damage my vehicle or other vehicles and I do not wish to kill other people." We can try to reach yet further intelligibility of her actions and ask, "why do you not wish to damage other people's vehicles or kill people?," and her reply will be, "because property and life are valuable."

We are now in a position to understand that citizen "c's" answers have a structure which is the structure of practical reason, where reasons are connected to other reasons, whose chain has a finality. The finality is provided by the agent from the first person or deliberative perspective when she advances a value or good-making characteristic that swallows the earlier stages of the action and provides intelligibility to the movements of "c's" body. This explanation seems primary and more fundamental than the explanation in terms of acceptance-beliefs, reason-beliefs as a

mental state of either primary or secondary rules of the legal system or exclusionary reasons.[14]

If citizen "c" decides not to stop at the red traffic light because she is driving her neighbour to the hospital, who is dying, then to the question "why are you not stopping at the red traffic light?," she might reply, "don't you see it? My neighbour is dying and I need to get to the hospital as soon as possible." And to the question, "why do you need to bring him to the hospital as soon as possible?" she might reply, "because I want to save his life"; in response to the question "why do you want to save his life?," the answer will be, "because life is valuable." This set of answers will give intelligibility to her actions, which includes the movements of her body and what she produces, i.e. a vehicle moving in the direction of the hospital, and will also explain why she did not stop at red traffic lights. Thus, she went through the red traffic light not because of her belief that on this occasion there was no valid legal rule, nor because of her belief that the rule of "stopping at red traffic lights" does not protect or ensure values such as property or life. Her mistake lies, arguably, in not "perceiving" that the life of her neighbour is as valuable as the lives of pedestrians and the drivers of other vehicles. Her mistake lies in her understanding of the goods or values at conflict in the particular situation.

The classical model of practical reasoning and intentional action laid out the view that for an action to be controlled and guided by the agent the reasons need to be in the action and therefore transparent to the agent (see Sect. 2, c). The answers to the question "Why"? provide the order of reasons that guarantees successful compliance with the legal rules and directives by the agent. They are the reasons in action that the agent has together with the values or good-making characteristics that the legislator and or judges aim to promote and want the citizens to "grasp" as the grounding of their actions. The transparency condition of practical reason warrants that the citizen is able to engage with the good-making characteristics that ground legal rules. But if the order of reasons is opaque, how can there be an action as intended by the legislator or judge as an order of reasons that has as a finality a value or good-making characteristics? If the reasons are opaque and you do something "because someone says so" you do not know "why" you are performing the action and therefore the action is not intentional. Furthermore, one might assert, the legislator, judge or official is not the origin of change and the origins of change are in external empirical factors, e.g. the fear mechanism that acts within the agent, psychological processes in the agent, mental states such as beliefs, acceptance-belief or reasons-belief.

References

Anscombe, E. 1957. *Intention*. Cambridge, Mass: Harvard University Press.

Anscombe, E. 1981. On the source of authority of the state. In *Ethics, religion and politics: collected philosophical papers*, ed. E. Anscombe. Hoboken, N.J.: Wiley-Blackwell.

[14]Raz's exclusionary reasons account (Raz 1999) privileges the theoretical point of view. See also note 2 above.

Aquinas. 2006. *Summa theologiæ*, trans. T. Gilby. Cambridge: Cambridge University Press.
Aristotle. 1934. *Nicomachean ethics,* trans. H. Rackham. Cambridge, Mass: Harvard University Press.
Aristotle. 1983. *Physics books III and IV*, trans. E. Hussey. Oxford Clarendon Press.
Aristotle. 2006. *Metaphysics book Θ,* trans. S. Makin Introduction and commentary. Oxford: Oxford University Press.
Chisholm, R. 1976. Freedom and action. In *Freedom and determinism*, ed. K. Lehrer. New York, N.Y.: Random House.
Coope, U. 2009. Change and its relation to actuality and potentiality. In *A companion to Aristotle*, ed. G. Anagnostopoulos. Hoboken, N. J.: Wiley-Blackwell.
Davidson, D. 1980. *Essays on actions and events*. Oxford: Oxford University Press.
Enoch, D. 2011. *Taking morality serious: A defense of robust realism*. Oxford: Oxford University Press.
Edgeley, R. 1969. *Reason in theory and practice*. London: Cornerstone-Hutchinson.
Evans, G. 1982. *The varieties of reference*. Oxford: Oxford University Press.
Finnis, J. 1998. *Aquinas*. Oxford: Oxford University Press.
Frede, M. 1994. Aristotle's notion of potentiality in metaphysics Θ. In *Unity, identity and explanation in Aristotle's metaphysics*, ed. T. Scaltsas, D. Charles, and M. Gill. Oxford: Clarendon Press.
Grisez, G. 1969. The first principle of practical reason, a commentary on the summa theologiae, 1–2, Question 94, article 2. In *Aquinas. A collection of critical essays*, ed. A. Kenny. London: MacMillan.
Hart, H.L.A. 2012. *The concept of law.* Oxford: Clarendon Press (1st ed. 1961).
Kenny, A. 1979. *Aristotle's theory of the will*. London: Duckworth.
Kosman, L.A. 1969. Aristotle's definition of motion. *Phronesis* 14: 40–62.
Lewis, D. 1969. *Convention*. Cambridge, Mass.: Harvard University Press.
Marmor, A. 2007. Deep conventions. *Philosophy and Phenomenological Research* 74: 586–610.
Moran, R. 2001. *Authority and estrangement*. Princeton, N.J.: Princeton University Press.
Pasnau, R. 2002. *Thomas aquinas on human nature*. Cambridge: Cambridge University Press.
Raz, J. 1979. *The authority of law*. Oxford: Oxford University Press.
Raz, J. 1986. *The morality of freedom*. Oxford: Oxford University Press.
Raz, J. 1999. *Practical reason and norms*. Oxford: Oxford University Press.
Rodriguez-Blanco, V. 2003. Is finnis wrong? *Legal Theory* 13: 257–283.
Rodriguez-Blanco, V. 2012. If you cannot help being committed to it, then it exists: A defence of robust realism in law. *Oxford Journal of Legal Studies* 32: 823–841.
Rodriguez-Blanco, V. 2014a. Does practical reason need interpretation? *Ragion Practica* 317–340.
Rodriguez-Blanco, V. 2014b. *Law and authority under the guise of the good*. Oxford: Hart-Bloomsbury.
Rodriguez-Blanco, V. 2016. Re-examining deep conventions: practical reason and forward-looking agency. In *Metaphilosophy of law*, ed. T. Gizbert-Studnicki. Oxford: Hart-Bloomsbury.
Rodriguez-Blanco, V. 2017. Practical reason in the context of law: What kind of mistake does a citizen make when she does violate legal rules? In *Cambridge companion to natural law jurisprudence*, ed. R. George, and G. Duke. Cambridge: Cambridge University Press.
Ross, W.D. 1995. *Aristotle's physics: a revised text with introduction and commentary*. Oxford: Oxford University Press.
Vogler, C. 2007. Modern moral philosophy again: isolating the promulgation problem. *Proceedings of the Aristotelian Society* 106: 345–362.
Wittgenstein, L. 1953. *Philosophical investigations,* trans. E. Anscombe. Oxford: Blackwell.

Part II
Kinds of Reasoning and the Law