

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION



VOLUME 5.2 /2025

FICTIONAL MINDS. THE LAW'S MISREPRESENTATION OF HUMAN THOUGHT

§§§

MICHELE UBERTONE AND GIUSEPPE ROCCHÈ (EDS.)

Table of Contents

Foreword I-XVII

ARTICLES

G. Peluso Lopes, *Debiasing Strategies and Judicial Decision-Making: Exploring a Duty to Improve Judges' Capabilities* 1-45

P. Capriati, *Fictional Mechanical Minds* 46-70

F. Scuderi Di Miceli, *When the Nudge Fails: The Limits of Behavioural Individualism and the Case for Meta-nudge* 71-106

M. Taroni, *Law and Surveillance in the Digital Age: The Role of Orientation* 107-141

M. Fernández Núñez, *Paternalistic Interventions: What do They Presuppose About Human Rationality, and When are They Justified?* 142-185

A. Artosi, *Can (Should) Law Do Without Free Will?* 186-203

CONFERENCE PAPERS

S. M. Cafaro, *Peace, War, Democracy: The Value of Different Perspectives* 204-209

A. Bojinović Fenko and J. Brsakoska Bazerkoska, *European Union's Actorness Amid the Weakening Liberal International Order in the Fields of Trade, Digital Sovereignty and Conflict Resolution* 210-239

E. Akgemci, *Lessons from Feminist Foreign Policies: Rethinking the EU's Role in Promoting Peace and Human Rights Amid Anti-Gender Politics and Authoritarian Populism* 240-288

ISSN

2724-6299

(ONLINE) <https://doi.org/10.60923/issn.2724-6299/v5-n2-2025>

Ownership

Alma Mater Research
Institute for Human-
Centered Artificial
Intelligence (Alma
Human AI)
Alma Mater Studiorum
Università di Bologna
Via Galliera, 3 40121
Bologna (Italy)

Publisher

Alma Mater Studiorum –
Università di Bologna
Alma Diamond – open scholarly
communication
Via Zamboni, 33 40126
Bologna (Italy)



If not differently stated in the Volume, all materials are published under the Creative Commons Attribution 4.0 International License.

Copyright © 2025, the Authors

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

EDITORIAL BOARD

DIRECTORS' BOARD

EDITOR-IN-CHIEF: Gustavo Gozzi (University of Bologna)

Alberto Artosi (University of Bologna), Giorgio Bongiovanni (University of Bologna), Susanna Maria Cafaro (University of Salento), Orsetta Giolo (University of Ferrara), Geminello Preterossi (University of Salerno)

ASSOCIATE EDITORS

Roberto Brigati (University of Bologna), Massimo Fichera (Maastricht University), Annalisa Furia (University of Bologna), Matteo Galletti (University of Florence), Claudio Novelli (University of Bologna), Raimondello Orsini (University of Bologna), Chiara Valentini (University of Bologna), Annalisa Verza (University of Bologna), Giorgio Volpe (University of Bologna), Michele Ubertone (Maastricht University)

ASSISTANT EDITORS

José Antonio Castillo Parilla (University of Granada), Riccardo Fornasari (University of Paris-Nanterre), Aytekin Kaan Kurtul (University of Huddersfield), Micol Pignataro (University of Bologna), Francesco Rizzi Brignoli (University of Bologna), Marc Shucksmith-Wesley (University of Huddersfield), Andrea Antonino Silipigni (University of Lisbon), Marta Taroni (University of Chieti-Pescara), Martino Tognocchi (University of Pavia), Iryna Zhyrun (Higher School of Economics)

MANAGING EDITOR

Luigi Sammartino (University of Milan)

EXECUTIVE EDITOR

Francesco Cavinato (University of Bologna)

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

SCIENTIFIC COMMITTEE

Robert Alexy (University of Kiel), Carla Bagnoli (University of Modena and Reggio Emilia), Francesco Belvisi (University of Modena and Reggio Emilia), Yadh Ben Achour (United Nations – Human Rights Committee), Damiano Canale (Bocconi University), Rossana Deplano (University of Leicester), Pasquale De Sena (University of Palermo), Alessandra Di Martino (University of Rome “Sapienza”), Javier Espinosa de los Monteros Sanchez (University of Anahuac), Matthias Klatt (University of Graz), Josep Joan Moreso (Pompeu Fabra University – Barcelona), Andrea Morrone (University of Bologna), Luigi Nuzzo (University of Salento), Baldassare Pastore (University of Ferrara), Giorgio Pino (University of Roma Tre), Corrado Roversi (University of Bologna), Alessandro Serpe (University of Chieti-Pescara)

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

VOLUME 5.2
2025

Table of Contents

Foreword I-XVII

ARTICLES

G. Peluso Lopes, *Debiasing Strategies and Judicial Decision-Making: Exploring a Duty to Improve Judges' Capabilities* 1-45

P. Capriati, *Fictional Mechanical Minds* 46-70

F. Scuderi Di Miceli, *When the Nudge Fails: The Limits of Behavioural Individualism and the Case for Meta-nudge* 71-106

M. Taroni, *Law and Surveillance in the Digital Age: The Role of Orientation* 107-141

M. Fernández Núñez, *Paternalistic Interventions: What do They Presuppose About Human Rationality, and When are They Justified?* 142-185

A. Artosi, *Can (Should) Law Do Without Free Will?* 186-203

CONFERENCE PAPERS

S. M. Cafaro, *Peace, War, Democracy: The Value of Different Perspectives* 204-209

A. Bojinović Fenko and J. Brsakoska Bazerkoska, *European Union's Actorness Amid the Weakening Liberal International Order in the Fields of Trade, Digital Sovereignty and Conflict Resolution* 210-239


E. Akgemci, *Lessons from Feminist Foreign Policies: Rethinking the EU's Role in Promoting Peace and Human Rights Amid Anti-Gender Politics and Authoritarian Populism* 240-288

Fictional Minds. The Law's Misrepresentation of Human Thought

MICHELE UBERTONE

Assistant Professor of Philosophy of Law, Maastricht University (Netherlands)


✉ michele.ubertone@maastrichtuniversity.nl

 <https://orcid.org/0000-0002-7304-3701>

GIUSEPPE ROCCHÈ

Postdoctoral Researcher of Philosophy of Law, Maastricht University (Netherlands)

✉ giuseppe.rocche@unipa.it

 <https://orcid.org/0000-0002-6426-1942>

1. Fictional Minds

The purpose of this introduction is to identify the topic of the editorial section¹. In the first part, we will try to explain in what sense legal practice may be said to rely on false presuppositions about the human mind. We will also consider some doubts readers may have about the legal relevance of examining such presuppositions and offer some reasons for treating them as important to understand the way in which legal institutions do and should operate. In the second part, we will consider the content of the different papers that make up the issue, analysing the similarities and differences in the authors' focus and approach.

According to Hans Kelsen, legal sciences differ from natural sciences not so much in the portion of reality they examine, but rather in the interpretative principle they adopt to examine it (Kelsen, 1950; 1991, 24). Any fact of the world can become legally relevant if only a community decides to *treat it as*

¹ Although both sections of this foreword were conceived and discussed jointly by the two authors, in order to comply with a requirement of Italian law, we declare that Michele Ubertone is the author of the first section and Giuseppe Rocchè the author of the second section.

legally relevant. What constitutes the specific scope of legal enquiry is this *attribution of relevance*, rather than the accurate *description* of empirical facts. Natural sciences observe phenomena to explain them in terms of causes and effects, according to what Kelsen calls the *principle of causality*. Law, on the other hand, when faced with the same phenomena, does not primarily consider causes and effects but rather presupposes the obtaining of empirical facts to then *impute* responsibilities and normative consequences: who is responsible for what and what must follow from a certain fact according to the legal norms in force. It relies on what Kelsen calls the *principle of imputation*. While a scientist might inquire whether a homicide *was* caused by drug use or genetic factors, a judge or a lawyer will consider whether and under what conditions that person *should* be regarded as responsible and whether the act should be punished, according to predetermined criteria.

Both causation and imputation establish nomological (law-like) relations among facts which can be used for drawing inferences, in the empirical sciences and in law, respectively. But there are important disanalogies between these two types of analyses. First of all, of course, scientific laws describe, while legal norms prescribe. Second, while causal analysis is virtually never-ending (any cause is itself caused and we may sensibly ask ourselves what caused the Big Bang, if anything), analysis based on imputation always has an endpoint, the attribution of responsibility to a person is warranted as long as it satisfies predetermined and authoritative criteria. Whereas science regards humans as objects determined by the physical world, the law assumes them as being capable of responding to reasons and considers them responsible in case of failure to do so. Kelsen makes clear that this conceptual operation – carving out certain elements of reality, legal agents, to remove them from the domain of causality and place them in that of imputation – is not committed to the existence of aspects of human behaviour which are exempted from causal laws.

If human behaviour, to be a possible object of imputation, would
have to be considered as exempted from the law of causality,

causality and freedom would be, indeed, incompatible. (...) However, there is no such conflict if we understand the true meaning of the statement that man as a moral, religious, or legal person is free (Kelsen 1950, 8).

In a Neo-Kantian vein, Kelsen conceives causality and imputation not as intrinsic features of reality but rather as two alternative *a priori* concepts we employ in order to understand reality (Paulson, 2001, 47).

According to Kelsen, in other words, assuming human responsiveness to legal reasons is a condition of possibility for legal discourse. This assumption is a transcendental matter, not subject to being disproved by scientific progress in the understanding of the actual *causes* of human decision-making. Saying that John's act of disobeying a legal norm can be fully explained by social or biological factors is not, in itself, an argument for claiming that John is not legally responsible for his action. The fact that John's behaviour is caused by something other and additional to his conscious choice does not exclude the possibility that such behaviour can nonetheless be attributed to him. This is because, whatever criterion of imputation a given legal system adopts, it must ultimately rest on normative criteria, criteria that are established by what Kelsen calls "acts of will" (Kelsen, 1950, 2). Once those authoritative criteria are satisfied, any examination of the causal origins of the behaviour becomes irrelevant. This line of reasoning – akin to other classical arguments, such as those derived from Hume's law – has operated as a protective barrier insulating legal theory from the growing influence of the cognitive sciences. As a result, legal theorists have often treated the findings of these disciplines as largely irrelevant to legal reasoning. Canale and Tuzet, in discussing the criticisms made on psychological grounds to the syllogistic model of legal reasoning, for example, write:

Sometimes the critic of the syllogism adds that it is merely an *ex post* rationalization of a decision already made on other grounds; this would occur especially in those legal systems that allow the

operative part of a judgment to be announced first and the reasoning to be drafted afterwards. Yet the reply remains that the objection misses the point: from the standpoint of what justifies a decision, what matters is the logical structure and the reasons set out in the grounds; these are the elements that will be assessed and on which any appeals must be based. The rest, including whatever runs through the judge's mind, is legally irrelevant (Canale and Tuzet, 2019, 23, our translation).

However, even if we accept the Kelsenian point that *some* kinds of facts about human behaviour must be disregarded to think legally at all, and the more general point that law is concerned with how legal authority *should* affect human behaviour rather than how it *does* affect it, this does not necessarily imply that it is rational for us always discard all discoveries about the human mind as irrelevant. We may well have good reasons to revise our criteria of imputation if we realise that they make legal practice unfit to pursue the results that we aim to achieve with it. Law is an activity that, like cooking, architecture, or medicine, human communities undertake in order to achieve certain results. The resolution of disputes, the coordination of large numbers of individuals, the repression and reduction of antisocial behaviour are primary goods, we typically aim to secure through law, just like nutrition, shelter, and health are goods we aim to secure through cooking, architecture and medicine, respectively. If these activities are carried out based on false factual assumptions, they risk producing effects other than those expected. In this perspective, it is not at all obvious that psychological facts should be disregarded when we think about how laws should be drafted or how they should be interpreted.

With this editorial proposal of *Athena*, we aim to explore some of the assumptions about how the human mind works that seem to underlie various legal institutions – assumptions that cognitive sciences are progressively challenging. In every contemporary democracy, it is assumed that citizens choose their representatives “freely and consciously”. Parliament translates

this presumed “popular will” into legal norms, which are then published so that anyone may “know” their content. Judges, in turn, are called upon to “decide impartially according to the law”, thereby implementing the collective will in concrete cases. Criminal sanctions are directed exclusively at those who have “voluntarily” broken the law, “knowing” in advance what the legal consequences of their actions would be. All these doctrines form pillars of modern legal culture, but they rest on highly problematic assumptions regarding how the cognitive processes of the actors involved in the legal system actually work. Can lawyers truly afford to disregard the possible falsity of these assumptions?

Here are some arguments that might suggest a positive answer:

1. A critic might say that law is not a set of ordinary speech acts and cannot be said to have presuppositions in the same way as ordinary speech acts do. Legal acts seem to be somewhat detached from the specific person performing them and their psychological states. Law consists of *authoritative* acts endowed with validity regardless of whether they are justified or conceived to be justified by those who issue them. This seems to be an essential aspect of the nomodynamic character of a legal system (i.e., its ability to regulate its own transformation): legal norms are understood as claiming authority for the mere fact of having been produced in a certain manner. A statute, a regulation, or a judgment does not condition its binding force on its *content*, but only on its *source*. They are valid insofar as they come from authorities legitimized to issue them according to certain procedures. This seems to be an essential component of legal practice. If the addressees of legal texts were always allowed to scrutinise the correctness of factual assumptions underlying legal directives and disregard them in case these turn out to be false, then law would lose its authority, and thus its main social function.
2. Even if law had presuppositions, the critic may argue, they would not concern the psychological underpinnings of human behaviour. Law is

concerned with external and observable behaviours and must necessarily disregard what happens in the inner forum of each individual subject. Again, this seems to be connected to the need for legal acts to limit possible disagreement about their content. As psychological facts are simply too hard to prove, it would be unreasonable for legislators and policy makers to condition their directives implicitly or explicitly to the obtaining of such facts.

3. Even if law did have presuppositions of this kind, in particular presuppositions about the rationality and autonomy of human beings, these presuppositions should not be too readily dismissed. Cognitive sciences may have challenged the idea of a fully rational and autonomous subject, but they have not provided conclusive evidence of the necessary irrationality of human beings or of the nonexistence of free will. Many experimental results are contested, partial, or valid only in specific contexts. As long as this conception is not definitively disproved by science, the legal system can continue to rely on it.
4. Finally, even if law did contain false presuppositions about the human mind, their falsity would not necessarily constitute a problem. Law can, and in some cases must, introduce *fictions*: representations that are not intended to reflect empirical reality, but rather to ensure the coherence and functionality of the legal order. What matters, from a legal point of view, is the ability of these fictions to support effective institutional practices, not their truth.

On the other hand, the following considerations may be made in response to this imaginary critic:

1. Being subject to the authority of law means having reasons for acting based on its commands rather than based on our own assessment of what would be best to do. This is what must be taken into account to secure the law's social function. However, for lawmakers to really provide such reasons and thus for law to have authority in this sense, according to Joseph Raz, one condition must necessarily be met.

Legislators must base their directives on first-order reasons that would apply to the people even independent of the directive itself (Raz 1986). In this perspective, when this condition is not met, in a particular case, the source of directives may lack authoritative power for that case, despite being generally authoritative. Although epistemic and practical reasons may sometimes behave differently, the mechanism of exclusion in the two cases is similar and thus a non-legal example can help illustrate the phenomenon we are interested in. Consider the case of a physician whom you regard as a legitimate authority and whom you consult to determine whether you suffer from a particular disease. She issues a diagnosis because of the symptoms you reported, and you treat her judgment as authoritative. However, suppose that after leaving the clinic a new symptom appears. While the physician's diagnosis remains authoritative – and this prevents you from second-guessing the disease the earlier symptoms may reasonably indicate – the new symptom constitutes an additional reason that cannot plausibly be excluded by the authority of the initial diagnosis. In general, even if we take legal directives to provide content-independent reasons (Hart 1982, 253–55), their status as practical reasons depends on their justification, that is, on their correct connection to the world, and thus on the truth of their presuppositions. This does not mean that individuals are *always* entitled to scrutinize the correctness of legal presuppositions and disobey whenever they believe them to be false. But it is not uncommon for lawyers to consider considerations based on whether a particular application of a rule fulfils its intended purpose. One of the reasons that may generate this situation is precisely the falsity of presuppositions.

2. The law routinely makes its effects depend on psychological facts. One need only think of *mens rea* in criminal law or of the meeting of minds in contract law. What generates legal uncertainty is not the mere requirement that such psychological facts be established, but rather

that the law often demands their subsumption under concepts that are difficult to operationalize, concepts articulated in folk-psychological terms that do not correspond to the vocabulary of the relevant sciences. On the other hand, the advances in cognitive science that prompt us to reconsider certain assumptions of legal practice are empirical advances. In this sense, the worry that relying on them would make the law indeterminate because they are too open to dispute seems unwarranted.

3. Even if cognitive sciences have not conclusively shown that human beings are entirely irrational or non-autonomous, they remain the best source for determining if and when this is the case. The belief that a certain set of presuppositions of a norm is *certainly* false may, under certain conditions, justify the belief that the norm will certainly fail to produce the intended effects. Likewise, the belief that a certain set of presuppositions of a norm is *probably* false may, under certain conditions, justify the belief that the norm will probably fail to produce the intended effects. If we aim to increase the instrumental rationality of legal practice, it is reasonable to take into account provisional and less-than-certain conclusions (as of course full certainty is never available in science).
4. Finally, false presuppositions in legal norms or institutional practices must be distinguished from legal fictions. When law intentionally introduces fictions, it does so because it is useful to treat a false proposition as true. For example, when several people die as a result of the same event and the exact time of death of each is unknown, the law presumes a fact that is almost certainly false: that they died at exactly the same time. But treating this false proposition as true produces positive and intentionally sought effects, such as legal certainty and the functionality of inheritance regulation. When law rests on false presuppositions that do not help achieve its objectives but instead hinder them, on the other hand, we are dealing with a very

different phenomenon. The analysis of false presuppositions is precisely aimed at distinguishing useful fictions from mere scientific ignorance in legal practice.

2. Structure and Key Terms of this Section

The papers collected in this issue address the problem of false presuppositions of legal practice about the functioning of human minds from different viewpoints.

Peluso Lopes's paper offers a critical examination of what debiasing strategies can enhance the rationality of judicial decision-making, particularly in light of the limitations of standard procedural safeguards.

Capriati's paper focuses on the rationality of decision-makers in political systems and, drawing on an epistemic conception of legitimacy, distinguishes different types of government, based on the different subjects whose rationality must be assumed to consider the government legitimate.

Scuderi's paper challenges a distinction that has emerged in the literature concerning *nudges*: that between behaviouristic social interventions targeting individuals (*i-frames*) and more traditional measures aimed at transforming institutional contexts (*s-frames*), engaging with the notion of *meta-nudge*.

Taroni's paper addresses the phenomenon of surveillance capitalism in the age of tech giants and develops two markedly different perspectives for confronting its threats.

Fernández Núñez's paper considers paternalistic interventions, identifying the various features that may justify restrictions on individual freedom, and critiques the legal use of general anthropological conceptions.

Finally, Artosi's paper is centred on free will in the context of Kelsen's theory of law, focusing on how the great philosopher defended a form of compatibilism.

Different as they might seem, the papers all deal with our problem in one way or another, and they can thus be analysed according to the following

three parameters: the *content* of the presupposition they deal with; the *subject* about whose mind the presupposition is made; and finally, the *subject* who makes this presupposition. Let's examine these three perspectives one by one.

Mental presuppositions have a certain *content* – that is, they adopt a particular image of how the human mind functions. A recurring target of many contributions is the overly optimistic representation of the human mind. Peluso Lopes discusses how legislators assume judges to be more rational than they actually are and tries to examine how the situation could be improved with particular debiasing strategies. In her view, traditional procedural safeguards are often problematic because they are compromised by a rationalist conception of legal judgment centred on the idea of the *legal syllogism*. The legal syllogism “places a general and abstract norm as a major premise; a representation of the fact as a minor premise; and draws a deductive conclusion, that is, the conclusion by a particular and concrete norm”. It is worth noting that although this image is regarded by the author as a *normative* model rather than a description of judges' cognitive activity, it remains true that for a legal system to adopt such a normative standard, it must be committed to the idea that the model can be realistically met by judges and other legal decision-makers. Yet such hopes are ultimately doomed, because

one of the objections that have historically been raised against the syllogistic model is that judges do not decide according to it but do so under the influence of other dynamics like their emotions, idiosyncrasies and preferences of various kinds. According to this view, the syllogistic reasoning would therefore only be an *ex post* rationalization of a decision already made on other bases.

In a similar vein, the papers by Scuderi and Taroni on the use of *nudge* also engage with the limits of human rationality. While for Peluso Lopes the problem lies in the incapacity of legal decision-makers to effectively shield their judgments from implicit biases, the problem addressed by Scuderi and

Taroni is that individuals are unable to process all the information relevant to the case at hand. In Peluso's case, the problem seems to be that humans are not able to disregard certain types of information to reach decisions in the way we would sometimes expect them to do; in Scuderi and Taroni, the problem is that humans are often unable to process all relevant data in the way we would want them to. More specifically, in Scuderi, the problematic assumption under examination consists in seeing human rationality as "an optimising choice process, in which individuals select the alternative that maximises their welfare, based on complete information and unlimited computational capacity". Similarly, for Taroni, classical liberal theory presupposes a model agent "who has emancipated himself from the 'state of minority', standing free among others and equal [...] both rational and reasonable, capable of articulating and pursuing their own comprehensive life-plans", a *homo oeconomicus* who knows what he wants and what are the most expedient means to pursue it. On these premises, the role of institutions is reduced to a supposedly neutral protection of individual self-determination. But again, such depictions of the human mind are *fictitious*: human reasoning must be conceived as an activity under constraints (as Herbert Simon made clear), and people's autonomy is a dangerous dogma (a recurrent theme in post-liberal theory).

Capriati's paper also concerns problematic presuppositions about human rationality but does so in discussing the legitimacy of political systems. Capriati considers *epistemic conceptions of legitimacy* and argues that, in such conceptions, a necessary condition of legitimacy consists in assuming the instrumental rationality of the actors endowed with decision-making power. Capriati does not dwell on the reasons why certain presuppositions are not realistic. But he insists that if we believe that there are *correct decisions* in political matters, and that political power is justified insofar as it serves as a tool to reach these, we are compelled to assume someone's rationality. We may disagree on *whose* rationality to assume, but we cannot avoid assuming it altogether.

Fernández Núñez's contribution focuses on the breadth of the assumptions that the law needs to make about the human mind. The author's aim is to question the idea that relying on broad anthropological assumptions might be necessary for both the justification and the critique of paternalistic interventions. He shows how in heated debates on fundamental rights, opposing assumptions do emerge: there are *Pelagians*, who trust people's discernment, and *Augustinians*, who are enamoured of the idea that people are "vulnerable, wounded, torn, ruined". The former (like Maniaci) are overly restrictive towards paternalism, while the latter (like Moreso) are overly complacent towards it. The author invokes the adoption of a more nuanced approach.

The conflict between two opposite presuppositions seems inescapable in Kelsen's view, as depicted by Artosi, in his paper on free will and responsibility. According to Kelsen "only man's freedom (that is, the fact that he is not subject to the law of causality) makes responsibility (and that means: imputation) possible"; on the other hand "The establishment of a normative, behaviour-regulating order, which is the only basis of imputation, presupposes that man's will is causally determinable, therefore not free". There seems to be no room for manoeuvre to escape from the impasse, as the general functioning of law seems to require two absolute, broad, and incompatible presuppositions. The solution is found then not in the abandonment of general presuppositions in favour of more circumscribed ones, but rather in the adoption of compatibilism, to the extent that, at the end, imputation does not require either the fact or the fiction of free will.

As is clear from the discussion above, the presuppositions considered in the different contributions concern the minds of different actors within the legal system. This is our second parameter of comparison.

Peluso Lopes focuses on judges; Taroni and Fernández Núñez discuss assumptions related to ordinary citizens. Ordinary citizens are also the focus of Scuderi's contribution, but he adds a further subject to the analysis: the *administrative machinery* that occupies the space, often neglected by legal

theory, between governors and the governed. Scuderi's thesis is that, although cognitive science has dispelled the rationalist illusion when it comes to private citizens, we still entertain a similar illusion regarding the activities of public administration. We continue to believe that while *nudge* is a suitable technique to guide private behaviour, interventions on the institutional structure, the so-called "s-frame", should be based on traditional instruments such as sanctions and economic incentives. Against this view, Scuderi argues that some contemporary interventions in administrative law can be better analysed in terms of *meta-nudges*, that is, nudges targeting public officials themselves. A defence or critique of such measures goes hand in hand with a defence or critique of the underlying assumptions of rationality.

For Capriati, the identity of the people whose mental characteristics must be presupposed is central. In his view, different political conceptions of legitimate power can be distinguished on the basis of *whose* rationality they assume. More specifically, he examines aggregative democracy and argues that it presupposes individual rationality of single voters; then he turns to deliberative democracy and argues that it presupposes the rationality of groups or discursive practices; and finally explores the possibility of a *machine government* which would require assuming the rationality of machines.

In Artosi's paper, dedicated to the presupposition of free will rather than rationality, the subject of the presupposition is the person who realizes the condition of the Kelesian legal norm, and that ought to be punished and rewarded accordingly. It is the freedom and responsibility of this subject that must be reconceptualized to be rendered compatible with the principle of causality.

The third and last parameter we will consider is the source of the presupposition, that is, the subject *making* it.

In his analysis of the connection between rationality and legitimacy, Capriati considers the people affected by government decisions and possessing the power to influence them. In other words, according to him, different

conceptions of legitimate political power depend on which entities are believed to be rational by the *governed*. A peculiar feature of epistemic conceptions of democracy, in Capriati's analysis, lies in the coincidence between the people *making* the assumption and the people *whose* rationality is assumed. This circularity is an epistemic corollary of the idea that, in democracy, power belongs to the people.

In Peluso Lopes's analysis, the sources of the presupposition are the *drafters of the procedural system*. They endeavoured to counter certain forms of prejudices: "there are several elements of substantive and procedural law that can be interpreted as sedimented debiasing strategies, serving to avoid legal decisions that are arbitrary, unfair or ill reasoned". But they neglected "to adequately address the negative effects of biases that operate beneath the level of consciousness, and which are not subject to direct introspection".

In Taroni's contribution, the presupposition lies instead in laws on consumer protection. In the era of surveillance capitalism,

[t]raditional frameworks of law and regulation, built on the assumption of autonomous and rational subjects, are increasingly inadequate when confronted with the pervasive and opaque use of behavioural influence techniques by private actors.

According to this view, one of the deepest problems of our time can be found in the fact that the agents of surveillance capitalism do not rely on fictional presuppositions and so they take advantage over political institutions that still do.

The assumption of free will seems to be a practical and theoretical necessity in Artosi's work, meant to make intelligible the principle of retribution. In this sense, the presupposition would be, in other words, an inescapable common heritage of our culture. Artosi makes the effort to show that we may get rid of the assumption instead, and that it is ultimately a matter of moral sensibility if we have strong resistance in punishing someone when we believe that she has been caused to do what she did.

The individuation of the authors of the presuppositions raises a further issue. Is it really so clear, as the previous discussion suggests, that institutions are permeated by presuppositions?

A distinctive feature of Fernández Núñez's contribution is his sceptical answer. When discussing paternalistic interventions, he observes that the law does not explicitly incorporate paternalism; rather, paternalism functions as a way to make sense of certain restrictions on autonomy. Yet many such restrictions are equally compatible with alternative justifications. As he notes,

[i]t is very easy for the same decision to be presented, either by the authority that promulgates it, that called upon to interpret it, or for the observer or critic, to identify, alongside paternalistic foundations, non-paternalistic foundations that are (or that are intended to be) complementary or alternative to the former.

Presuppositions, then, lie in the eyes of the beholder. It is often less the legislators than the legal scholars and philosophers who boldly ascribe a paternalistic foundation, and the accompanying assumptions of irrationality, in order to make sense of certain legal outcomes. But this interpretive attitude is questionable, as the *Lochner* and *Wackenheim* cases illustrate, and even dangerous, since it perpetuates the belief that our moral common sense is inherently paternalistic.

A similar scepticism towards the idea that legislators presuppose the agents' rationality can also be found in Scuderi. The popular distinction between the *i-frame* (concerned with individual behaviour and based on behavioural techniques like *nudge*) and the *s-frame* (dealing with traditional bureaucratic regulation) suggests that while in the *i-frame* the assumption of rationality has been replaced by a more realistic view of human behaviour, the regulation of bureaucracies still presupposes bureaucrats' rationality. Scuderi challenges this stark dichotomy. While such presuppositions of rationality are indeed present, they are neither necessary nor inevitable to make sense of certain legal interventions. As noted above, the author argues

that a recent development in Italian administrative law (the introduction of the *principle of mutual trust* between officials and citizens) can be seen as a *meta-nudge*, a nudge directed at the intermediaries who shape others' behaviour. Presuppositions, again, may lie more in the eyes of the scholars inclined beholder than in legal reality; yet they remain relevant, since law is a socially constructed entity inseparable from legal culture.

Before concluding, it is worth highlighting another common thread shared by several of the contributions. As we have seen, some of the authors in this issue take as their starting point the limits of human rationality and cognitive capacity, limits that are often overlooked by the institutional framework or by legal culture. Yet they do not stop at this rather discouraging diagnosis; instead, they move beyond it to develop a *pars construens*, a constructive line of analysis that seeks to build upon, rather than merely lament, both these limitations and the failure to adequately acknowledge them.

In particular, Peluso Lopes focuses on the measures that may be adopted to counter implicit bias in judicial decision-making. She advocates the introduction of auditing systems to make judges aware of unconscious influences on their reasoning, screening for cognitive decline, assessments of judges' physical and mental health, and debiasing training as an integral component of their continuing education.

Scuderi's recommendations, by contrast, do not target the legislator but rather legal doctrine, which tends to assume, unjustifiably, that "behavioural tools are inherently incapable of producing structural effects". The author's theoretical effort seeks to overcome this illusion, which he regards as a dangerous distraction from building effective systemic interventions.

Finally, turning to Taroni's paper, the reader will find not only a discussion of the importance of resisting certain fictional presuppositions but also a reflection on the various ways such resistance may take shape. On the one hand, Taroni contends that an effective means of countering the influence of the web giants of surveillance capitalism is to adopt "strategies analogous to those employed by the surveillance capitalists themselves", using

institutional nudges to counteract exploitative nudges. On the other hand, she draws attention to Stegmaier's *philosophy of orientation*, a distinct approach, rooted in a philosophical tradition and sensibility different from those explored by the other authors, centered on the awakening of individuals' rational capacities.

References

- Canale D. and Tuzet G. (2019). *La giustificazione della decisione giudiziale* (Giappichelli).
- Kelsen H. (1991). *General Theory of Norms* (Clarendon Press).
- Kelsen H. (1950). Causality and Imputation, in *Ethics*, vol. 61(1), 1-11.
- Hart H. L. A. (1982). Commands and Authoritative Legal Reasons, in H. L. A. Hart, *Essays on Bentham: Studies in Jurisprudence and Political Theory* (Clarendon Press), 243–268.
- Paulson S. (2001). Hans Kelsen's Doctrine of Imputation, in *Ratio Juris*, vol. 14(1), 47-63.
- Raz J. (1986). *The Morality of Freedom* (Oxford University Press).



ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

Debiasing Strategies and Judicial Decision-Making: Exploring a Duty to Improve Judges' Capabilities

GIOVANA PELUSO LOPES

*Postdoctoral Researcher, Tilburg Institute for Law, Technology, and Society (TILT)
Tilburg University (Netherlands)*

✉ g.lopes@tilburguniversity.edu

 <https://orcid.org/0000-0003-4798-2542>

ABSTRACT

Judicial decision-making carries profound consequences for individuals and society, with expectations that judges act with objectivity and neutrality. However, traditional legal frameworks assume a rationalist model of reasoning that inadequately addresses implicit biases. While existing legal safeguards target explicit instances of bias, they fail to account for biases that subtly influence judicial perceptions and judgments. This article explores how current legal norms, grounded in a syllogistic model of judicial reasoning, are insufficient to address implicit biases. Drawing on Behavioural Realism, the analysis demonstrates how judges remain susceptible to implicit biases despite their training and commitment to impartiality. The article proposes integrating debiasing interventions into judicial practice, including habit modification, environmental changes, and decision-support tools. It argues that such interventions should be encompassed within judges' professional duties, making their adoption mandatory under certain conditions.

Keywords: judicial decision-making, legal reasoning, implicit biases, judicial biases, debiasing

ATHENA

Volume 5.2/2025, pp. 1-45

Articles

ISSN 2724-6299 (Online)

<https://doi.org/10.60923/issn.2724-6299/22547>



1. Introduction

Every day, judges are tasked with rendering decisions that carry profound consequences both for the individuals directly affected by them and for society as a whole. For instance, they can decide whether a person should be sent to prison and for how long, as well as whether she can await her trial in freedom or have an early release. At the highest levels, judges have the power to interpret constitutional principles and to determine important social disputes involving fundamental rights. These decisions not only affect present cases, but can also reverberate to the future through the doctrine of precedent. Overall, judges hold great power in our society, and in wielding this power, we expect them to act with objectivity and neutrality. We expect judges to reach and justify their decisions based on authoritative legal sources, and to not be influenced by external factors that are irrelevant to the dispute.

There are many elements of substantive and procedural law that serve to fulfil these expectations, aiming to avoid legal decisions that are arbitrary, partial or ill reasoned. Examples include the right to appeal and the right to counsel, the right to reasoned verdicts, and the inadmissibility of evidence deemed prejudicial. While all of these aims to ensure accuracy and objectivity in judicial decision-making, they assume a normative model of legal reasoning that does not entirely reflect the reality of adjudication. According to this normative model, judicial decision-making follows a syllogistic and linear pattern: the judge begins with a specific set of facts, looks at the law that applies to those facts, and then reaches a verdict – all in a step-by-step, conscious, intentional and controllable process.

However, as representatives of the Legal Realism movement noted already a century ago, this model often fails to correspond to the reality of judicial practice, with extraneous factors such as judges' emotions, idiosyncrasies and preferences of various kinds having an impact on case outcomes. More

recently, some scholars started to advocate for a new type of Legal Realism, namely Behavioural Realism – a school of thought that asks the law to take into account more accurate models of human cognition and behaviour (Kang 2023, 5). This behaviourist strand of realism examines the extent to which factors that operate under the level of consciousness, such as implicit biases, influence the decision-making process. Judges, despite their training and commitment to impartiality, are not immune to the pervasive influence of implicit biases, and a growing body of research investigates how these can subtly shape their perceptions, judgments, and decision-making processes throughout various stages of legal proceedings.

This article explores how, by assuming a rationalist model of reasoning, the norms that govern the judicial process and judges' behaviour end up only addressing explicit instances of biased decision-making and are therefore insufficient to address the influence of implicit biases on judges' decisions. It argues that, in light of this limitation, alternative solutions to mitigate implicit bias in judicial decision-making are necessary. Here, it is proposed that judges' decision-making capabilities can be improved by certain interventions that alter their habits, modify the environment within which they interact, or provide tools to support decision-making. To the extent that these interventions seek to ameliorate bias, they are collectively referred to as debiasing interventions. Given the significance of judicial decision-making for litigants and society, the article explores the possibility of encompassing such debiasing interventions within the scope of the judicial professional duty of competence and diligence, thereby making their adoption mandatory under certain conditions.

Section 2 briefly presents the literature on implicit bias and the various ways in which judges are prone to various cognitive and social biases. It also addresses the different environmental and personal conditions that are conducive to biased reasoning. Section 3 explores how some strategies aimed at mitigating biased and ill-reasoned judicial decisions are already incorporated to a certain extent into procedural and substantial law, to ensure

its fair and impartial application. Notwithstanding, it argues that these measures are limited to explicit instances of bias and thereby do not fully prevent implicit biases from creeping into judicial decision-making. This is because they are aligned with a syllogistic normative model of judicial reasoning. Section 4 describes such model as well as the criticisms that have been addressed to it by the Legal Realism movement. It then explores a particular strand within this movement, namely, Behavioural Realism, along with its impact on the understanding of judicial decision-making. Section 5 further explores the concept of debiasing, presenting its main features and the different ways in which debiasing strategies can be classified. Consistent with the objectives of the Behavioural Realism movement, it discusses some commonly proposed debiasing strategies in light of their effectiveness in legal contexts, including the promotion of general bias awareness, training in rules and representations, the adoption of checklists, consider-the-opposite, exposure to stereotype-incongruent models and auditing. Finally, Section 6 provides general recommendations aiming at mitigating judicial implicit biases, framing them as part of judges' professional duty of competence and diligence. Section 7 concludes this article by summarising its main findings.

2. The Influence of Implicit Biases on Judicial Decision-Making

Before exploring the ways in which biases can affect judicial reasoning and decision-making, it is necessary to lend some precision to terms that are frequently encountered in the literature and to clarify the terminology that will be employed throughout this article. The first clarification relates to the broader distinction, commonly found in the cognitive and social psychology literature, between implicit and explicit biases (or, equivalently, unconscious and conscious biases). The first type is deemed implicit because it is generally considered to be latent, meaning that subjects tend to be unaware of them. The opposite is true for explicit biases, meaning that they are consciously accessible through introspection and endorsed by the individual. This

distinction is relevant because while there is a terminological conflation between the notion of explicit bias and prejudice, these fundamentally differ from implicit or unconscious bias – and this distinction is further reflected in the ways in which substantive and procedural law deal with judicial bias. In emphasising the impact of implicit bias, however, it is not suggested that explicit biases are unimportant or less significant to judicial decision-making. But the focus on implicit biases is justified in it being a category of bias that cannot be easily relegated to a few ‘bad apples’ or extremists, but that is highly pervasive even to individuals who are committed to ideals of justice and fairness. As Jerry Kang (2021, 81) rightly emphasizes, “implicit bias is here, right now, in your own courtroom, in your own mind, and in mine.”

The larger category of implicit bias can be divided into two main types, both of which have the potential to reduce the accuracy of a judgment: cognitive biases, which entail some broadly erroneous form of reasoning, and social biases, which entail reasoning based on implicit attitudes and stereotypes (Zenker 2021). The specialised literature distinguishes between cognitive biases and social biases, even though it can be argued that social biases also have a strong cognitive rooting. In this framework, the term cognitive bias is used within cognitive psychology – roughly equivalent to a cognitive fallacy – whereas social bias is used in social psychology to denote an unintentional partiality toward certain social groups. This distinction becomes more relevant when addressing the roots of bias (e.g., the fact that the majority of criminal defendants are black might reinforce judges’ implicit association between blackness and criminality) and how to mitigate it (e.g., social biases, being primarily a social problem, might require different forms of intervention, including ones that are more group focused). In the end, this work will deal with (ways to avoid/mitigate) both cognitive and social biases. Finally, note that social biases are sometimes also referred to as invidious or implicit biases. There is a risk, however, that adopting the latter definition may erroneously lead to the idea that cognitive biases are not also implicit (in

the sense of operating under the level of consciousness), which is why preference is given to the term social bias.

Implicit biases are inherent in human judgment, and judges are not excepted from their influence. Even if not consciously endorsed by judges, implicit biases can have a meaningful impact on the administration of justice, influencing judges' decisions throughout different stages of judicial proceedings: confirmation and hindsight bias, for instance, can hamper the hearing process, and racial bias can influence the way in which a witness or defendant is assessed. Likewise, the contrast/compromise effect, along with the gambler's fallacy or the status quo bias can affect judicial rulings, and sentencing can be anchored toward the initial demand made by the prosecutor or display gender or ingroup biases. To give some examples of relevant findings:

- Numerous studies have investigated the impact of anchoring effects on judicial decisions (for a recent meta-analysis, see Bystranowski et al. 2021). Experiments conducted with professional judges as subjects found that sentencing decisions and compensation awards were not only to be anchored by the initial demand made by the prosecutor, but also by random and unrelated factors to the decision at hand (Rachlinski et al. 2015; Guthrie et al. 2001, 2009; Englich et al. 2006; Enough and Mussweiler 2001).
- In a criminal investigation scenario, irrelevant contextual information affected judges' conviction rate, and confirmation bias led them to prefer incriminating investigations (Rassin 2020). Similarly, the pretrial detention of defendants later influenced judges' assessments of their guilt in criminal cases (Lidén et al. 2019).
- Judges' decisions were biased by the gender of the parties in studies involving hypothetical cases about child custody and relocation, employment discrimination, and criminal sentencing (Miller 2019; Rachlinski and Wistrich 2021).

- In negligence assessments, judicial decisions have also been found to be affected by hindsight bias, with judges in the hindsight condition believing that they would have foreseen the harm to a greater extent than those with genuine foresight did (Oeberst and Goeckenjan 2016).
- Data analysis of judges' bail decisions revealed racial bias against black defendants, even after controlling for variables such as criminal history and past pretrial misconduct (Arnold et al. 2020). And when tested with the Implicit Association Test (IAT), judges harboured the same measure of implicit biases concerning black people as most lay adults (Rachlinski et al. 2009).

The presence of bias is hard to separate from conditions that also produce suboptimal decision-making outcomes, or conditions that do so by interacting with a bias (Zenker 2021). These include environmental or institutional conditions, as well as personal constraints that, by generally leading individuals to be more reliant on intuitive thinking, are conducive to biased reasoning. Concerning the former, not only do judicial decisions often hinge on interpretations of subjective, ambiguous information, but the very nature of the trial process makes it particularly susceptible to implicit biases. For instance, it requires judges to make decisions about past events, opening the door to hindsight bias, and to convert assessments of harm into monetary values or prison terms, which is sensitive to anchoring and contrast effects. Unlike a jury, judges cannot be shielded from inadmissible evidence, even though there is no indication that they are better at disregarding it than jurors (Wistrich, Guthrie, and Rachlinski 2005). And due to their relevance, they cannot be shielded from certain information (e.g., identities or demographic characteristics) concerning litigants that could introduce bias into the decision-making process. Furthermore, the institutional context within which judges operate often lacks timely and constructive feedback, configuring a “wicked learning environment” (Hogarth 2010, 2015), and existing forms of accountability primarily concentrate on a judge's performance in individual

cases rather than conducting systematic long-term assessments that might reveal implicit biases (Leibovitch 2021).

Regarding personal constraints that interact with bias and produce suboptimal decision-making outcomes, fatigued judges may be more likely to rely on recommendations provided by surrounding court actors, such as the prosecutor, or to keep things as they are, defaulting to the status quo of, for example, maintaining incarceration (Danziger, Levav, and Avnaim-Pesso 2011). One of the manifestations of mental fatigue is decision fatigue, described as an impaired ability to make decisions and control behaviour as a consequence of repeated acts of decision-making, which can lead individuals to be “more susceptible to the use of cognitive heuristics (...) that may bias decision-making – potentially resulting in decisions that yield undesirable outcomes” (Pignatiello et al. 2020, 7). The effects of decision fatigue on judicial decisions have been documented (Shroff and Vamvourellis 2022; Torres and Williams 2022); as well as the effects of sleep deprivation – a situational antecedent implicated in the development of decision fatigue (Cho et al. 2017). Like most actors in the judicial system, judges are stressed, overburdened, and operating under time pressure. Significant stressors that judges face include the burden of consequential decision-making, exposure to disturbing evidence, and isolation (Maroney et al. 2023). Evidence suggests that stress leads people to scan alternatives less systematically and completely (Keinan 1987), and that time pressures are correlated with less accurate decisions (Braun 2000; Axt and Lai 2019), all of which is particularly problematic when it comes to judges making highly consequential decisions.

A final facilitating condition for judgment errors relates to the natural decline of cognitive capabilities due to aging. Empirical research shows, via an array of methodologies, the vulnerability of cognitive abilities like reasoning and processing speed due to the normal aging process. These findings are consistent across genders, educational levels, and generations (Kaufman et al. 2016). The declines in key cognitive abilities with age impact

individual executive functioning in areas such as judgment, decision-making, problem solving, concept formation, attention, memory, concentration, and planning ability (Flanagan et al. 2013). As Alan Kaufman (2021, 7) emphasizes, “this set of skills bears an intuitive relationship to the diverse kinds of real-life problem solving and effective processing of in-depth information required to be an effective, intelligent judge”. To the author, the overwhelming consistency in the literature on ageing and intellectual decline should call into question practices concerning the length of judicial appointments, considering that in many jurisdictions judges are, effectively, appointed for life, without checking potential cognitive decline.

It is important to acknowledge that empirical studies that directly investigate the effects of implicit biases on judges, using judges themselves as research subjects, remain relatively scarce. Nevertheless, existing evidence raises concerns: judges tend to perform no better than laypersons on standard measures of cognitive and social biases, such as the Cognitive Reflection Test (Guthrie et al. 2007) and the Implicit Association Test (Rachlinski et al. 2009). This suggests that they are not immune to the same cognitive limitations and automatic associations that affect decision-making more broadly. While the precise size of bias effects in judicial contexts is still debated, it should be noted that even minor disparities in treatment can accumulate over time, creating powerful headwinds and tailwinds that shape the trajectory of individuals in significant ways: when aggregated across entire groups – for example, between men and women or across racial categories – these small differences may generate substantial structural inequalities (Kang 2021). Moreover, for a single defendant, implicit biases can exert an influence at multiple points in the criminal justice process, from policing and charging decisions to bail, plea bargaining, pretrial motions, assessments of witness credibility, determinations of guilt and sentencing, and these accumulations can ultimately produce larger inequalities than those captured by individual studies measuring a specific bias alone.

Implicit biases are capable of exerting a significant influence on judicial decision-making across various stages of the trial process, being further aggravated by institutional structures, environmental conditions, and personal constraints. Yet, while recognising the pervasiveness and complexity of such biases, it remains necessary to assess the extent to which the law itself provides mechanisms to counteract them. The following section therefore turns to the limits of adjudicative debiasing through law, examining how existing substantive and procedural safeguards operate as tools of explicit bias mitigation, and why they may fall short in addressing judicial implicit biases.

3. Adjudicative Debiasing Through Law

Adjudicative debiasing strategies (Jolls and Sunstein 2006) aim to eliminate or mitigate biases in adjudication, largely focusing on procedural rules governing how judges or juries ought to decide. Indeed, there are several elements of substantive and procedural law that can be interpreted as sedimented debiasing strategies, serving to avoid legal decisions that are arbitrary, unfair or ill reasoned. In other words, they seek to avoid, or at least mitigate, negative decision-making outcomes to which cognitive and social biases can contribute.

Perhaps the most basic procedural debiasing measure rests in judges' accountability, insofar as decisions are subject to revision by higher courts (Arkes 1998). Article 2 of Protocol 7 of the European Convention of Human Rights (ECHR) provides that "everyone convicted of a criminal offense by a tribunal shall have the right to have his conviction or sentence reviewed by a higher tribunal (...)" (Council of Europe 1984). Mainly when it comes to decision-making at the higher level, having entire courts, rather than only individual judges deciding the case, can also be interpreted as a traditional and institutionalised debiasing measure (Zenker 2021). A related procedural form of intervention includes separating the decision-making process into

distinct phases, as is already the case in legal proceedings (Zenker, Dahlman, and Sarwar 2015, 3).

The debiasing technique known as consider-the-opposite (or playing the devil's advocate) operates by reminding agents of the hypothetical possibility of the opposite standpoint, requiring them to imagine alternate outcomes. In law, it is represented in the dictum *audi alteram partem*, which states that the other side must be heard as well, and enshrined in the principle that both parties should be given a fair hearing, and have the opportunity to respond to claims being brought against them. Such a principle is closely related to the right to counsel, considering how having a third party point out alternative plausible scenarios increases the likelihood of success of the technique:

(...) if whatever the opposite to consider is yet to be generated from imagination, rather than being made available in full detail by a defence attorney, for instance, then the same strategy may be less likely to work. Presumably, sheer lack of imagination, or lack of motivation to probe it, could lead agents to conclude (incorrectly) on the basis of the few alternatives that are thus generated – of which some may be ‘out of the way’ possibilities – that the original belief was not problematic after all. It is easy to see that the role of the defence attorney may be viewed as a counter-weight to a lack of imagination, for instance by providing plausible scenarios that put the defendant, at the relevant time, at places other than the crime scene, say, or explanations for the presence of the evidence presented by the prosecution even if the defendant was innocent (aka the probability of the evidence conditional on the falsity of the hypothesis) (Zenker, Dahlman, and Sarwar 2015, 17-8).

Another debiasing technique that corresponds to legal procedure is providing reasons in favour and against (*pro et contra*) a decision, which translates into the obligation, generally falling upon the judge, to give reasoned verdicts. In deciding cases at the domestic level, the courts in Council of Europe member states are usually obliged under domestic law to

provide detailed reasoning to their judgments. And the ECHR, as interpreted by the European Court of Human Rights (ECtHR) in its case law, also gives rise to substantial obligations concerning the content of that reasoning, primarily under Article 6 of the Convention. At the EU level, article 47 of the Charter of the Fundamental Rights (CFREU) guarantees the right to a reasoned judgment, also to be interpreted in accordance with the ECtHR case law.

A number of benefits theoretically achieved by giving reasoned judgments have been identified, which are summarised by Mathilde Cohen (2015) according to the three central values promoted by reason-giving in the courtroom, namely, participation, accountability, and accuracy. Fundamentally, imposing an obligation to give reasons for a judgment aims to secure litigants' involvement in the judicial process while ensuring the court is held accountable and accurate decisions are reached. To demonstrate that they are receptive to the evidence and arguments put out by the parties, judges must provide an explanation for their rulings. A judge conveys the degree to which the parties' arguments have been comprehended, accepted, or served as the foundation for the verdict by explaining their decisions. According to social psychology research, the perception that the decision maker has given due consideration to the respondent's views and arguments is crucial to accepting both the decision as well as the authority of the institution that imposes the decision (Cohen, Lind, and Tyler 1989). Giving reasoned verdicts also works as an accountability-enhancing mechanism. It limits judicial discretion by ensuring that written decisions or at least some record of the proceedings can be read and reviewed by higher courts, as well as the public in general, and by encouraging judges to treat similarly situated cases alike and to treat differently situated cases differently (Cohen 2015). Lastly, giving reasons for judgments enforces a form of self-discipline that is thought to improve the quality of the decisions themselves – a process often described as the 'it won't write' phenomenon:

In attempting to reason her decision, a judge discovers that she cannot find an appropriate legal justification, leading her to reconsider her initial ruling and make a more accurate determination. The theory can be generalized to non-written forms of justification: forcing judges to substantiate their decisions based on facts and legal arguments enhances the accuracy of judicial decision-making. It ensures that judicial decisions are not made arbitrarily or based on speculation, suspicion, or irrelevant information. The giving of reasons, it is thought, ensures that the deciding court has considered all relevant factors, researched the applicable law and given the case the thought it deserved (Cohen 2015, 511-2).

A final debiasing strategy is censorship, which corresponds to insulating the decision-maker from information that may lead him or her to biased decisions. According to the law of evidence, even if certain pieces of proof are factually relevant to the dispute, they may not be admitted in trial when, e.g., its potential to confuse or mislead jury members outweighs its probative value. Procedural rules also insulate judges from acting in certain situations, in order to fully implement the principle according to which every trial takes place before an independent and impartial judge. These are situations identified by the legislator where there is a risk of biased reasoning due, for instance, to the judge having an interest in the case assigned to him, or being in a relationship of kinship, affinity or marriage with one of the parties. In these cases, the judge vested with the decision of a case has the obligation to abstain from judging, and must be replaced by another. A risk of biased reasoning may also emerge if the judge has already carried out certain actions in the proceedings – for instance, if he or she pronounced or contributed to pronouncing the sentence in one level, exercising the functions of judge in the other levels of the proceedings is prohibited.

In short, it is possible to re-identify elements of modern procedural and substantive law as sedimented debiasing techniques. Nevertheless, biased reasoning and decision-making persists amongst judges, as illustrated by the

findings presented earlier in this article. A possible explanation for this is that formal procedures typically address explicit instances of biased decision-making. While the right to appeal serves the general goal of ensuring the correct interpretation and application of the law, it is a remedy often limited to specific situations. For instance, the grounds for cassation are limited to when there is a failure to properly apply the substantive law to the facts found true by the first instance court, or a failure to apply procedural rules that give rise to a nullity. And even though an appeal trial is supposed to address both the facts and the law by freshly retrying the case in its entirety, in most systems today appellants-to-be must articulate reasons for exercising this remedy, with courts being able to dismiss an appeal without a hearing if the reasons are considered inadequate. The decisions of courts of appeal are, in this sense, becoming increasingly “cassational” in that they review the judgment of the first instance court for factual and legal errors before admitting the case for trial (Thaman 2019). It is thus hard to envision how a litigant could successfully challenge a decision through one of these remedies on the grounds of it having been unduly influenced by implicit biases.

When it comes to insulating (or censorship), the law of evidence has developed with the main goal of keeping even relevant evidence away from a frequently distrusted jury, with no specialised training either in law or in factual analysis (Schauer 2009, 210), while holding the belief that judges, precisely because of their training, are better at assessing, weighing and, when necessary, ignoring evidence. While it may be possible to shield the jury from evidence deemed inadmissible, there are reasons to be sceptical of the effectiveness of this insulating technique when it comes to judges (see Wistrich et al. 2005).

Lastly, the duty to give reasoned verdicts seeks to ensure that judicial decisions are not made arbitrarily or based on irrelevant information, and there is some (albeit limited) evidence that suggests that requiring decision-makers to explain their reasoning for a decision may help mitigate some cognitive biases (see Guthrie, Rachlinski, and Wistrich 2007; Mitchell 2002;

Zenker et al. 2018; Liu 2018). According to traditional, formalist theories of judging, judicial decision-making follows a linear pattern: the judge searches for relevant evidence, weighs it and coordinates it with the applicable law and the dominant doctrines before reaching a decision – all in a step-by-step, conscious, intentional and controllable process. Once the decision is made, the judge who is asked to give a justification simply recounts these steps (Cohen 2015). However, the picture depicted by a number of psychologists offers a different account, suggesting that people's reasoning is frequently unconsciously motivated, in the sense that the reasoning process constructs *post hoc* justifications (see generally Haidt 2001, 2013).

In sum, existing substantive and procedural rules designed to prevent arbitrary, partial, or inaccurate legal decisions primarily focus on the explicit manifestations of biases, neglecting to adequately address the negative effects of biases that operate beneath the level of consciousness, and which are not subject to direct introspection. In order to effectively mitigate the effects of implicit biases, the law needs to take into account the influence of these cognitive and behavioural factors in judges' decision-making, and incorporate new measures aimed at identifying and counteracting their impact. The Legal Realism movement recognises the influence of such extraneous factors in judicial decision-making, challenging the notion of purely objective legal reasoning. By acknowledging that judicial decisions are shaped not only by legal rules but also by psychological and social dynamics, legal realism provides a more accurate account of legal reasoning – one that supports the implementation of targeted interventions to address implicit biases and promote greater fairness and impartiality in the legal process. The next section therefore turns to Behavioural Realism, a contemporary strand of the legal realist movement that builds upon these insights. Unlike formalist theories of judging, it explicitly incorporates unconscious influences on decision-making, such as implicit cognition and bias, into its account of legal reasoning.

4. Behavioural Realism: Accounting for Unconscious Influences in Legal Reasoning

Broadly understood, reasoning consists in the process of drawing conclusions (inferences) from some initial information (premises). One form of doing so is through deduction, a reasoning process in which a conclusion necessarily follows from the stated premises. The paradigm of deduction is syllogism, which consists of a general statement, known as the major premise; a specific statement, known as the minor premise; and a conclusion that necessarily follows from the two premises (Eisenberg 2022). A famous example of a major premise is that “all men are mortal”; as a minor premise, “Socrates is a man”; from that, one can deductively arrive at the conclusion that “Socrates is mortal”. In deductive legal reasoning, the decision-maker, most frequently a judge, begins with a specific set of facts, looks at the law that applies to those facts, and then reaches a verdict.

This type of legal reasoning is described by Damiano Canale and Giovanni Tuzet (2020) as the judicial syllogism, consisting in a prescriptive (rather than descriptive) model of judicial practice. It places a general and abstract norm as a major premise; a representation of the fact as a minor premise; and draws a deductive conclusion, that is, the conclusion by a particular and concrete norm. The internal justification, in this scenario, is the justification that the premises give to the conclusion of the syllogism (Canale and Tuzet 2020). When the facts of the case are proven in accordance with the relevant standards of proof, and they clearly fall under the range of application of the rule, the judge’s conclusion will follow deductively, justified by simple mention of the facts and the legal rule. And being a deductive conclusion, it is necessarily true if the premises are also true. The issue, however, is that in most cases the premises are not already formed, not even the legal ones, but must be figured out by the judges during and at the outcome of the trial, on the basis of the arguments produced by the parties and further considerations that are legitimate and relevant. They must determine the premises starting

from all the relevant evidence, legal texts and materials and derive the rule applicable to the case. There are many ways in which ambiguity can creep into this process:

For once, the decision-maker is faced with a specific set of facts. If he or she is a judge, there are almost always two versions of the facts. It is the attorneys' job to organise the facts in a way that fits the legal outcome they wish to achieve, and they do this by emphasising different facts and, often, different legal precedents (...) There may be more than one law that is potentially applicable. There may be several statutory provisions that might be relevant, and the two opposing counsels may argue that a different rule is the one that should control this case. The statute itself may violate a higher rule, such as the (...) constitution. The rule may be ambiguous, as in a ban on 'excessive noise', or the application of the 'reasonable person' standard ('Would a reasonable person have believed that her life was in danger?') (Ellsworth 2005, 686-7).

Therefore, further reasons are needed to establish the truth or correctness of the premises, which involves the elaboration of other arguments in addition to the syllogism. In other words, additional to the connection between the premises and conclusion of the syllogism, reasons are required to assume the soundness of both premises. Canale and Tuzet (2020) hence distinguish between the internal justification, which is the justification of the conclusion of the judicial syllogism, and the external justification, which is the justification of its premises. While the former is typically deductive, the latter is more often non-deductive. Within the external justification, the authors further distinguish between the external justification in law, which relates to the major premise of the syllogism, and the external justification in fact, which relates to the minor premise. Interpretative or integrative arguments of the law offer the external justification in law, which consists in drawing rules from legal provisions (interpretative arguments) and in filling any gaps in the

system (integrative arguments, the main which is the analogy). Meanwhile, offering the external justification in fact is evidentiary argumentation, consisting in drawing evidentiary conclusions from empirical evidence or other factual information collected in the trial or proceeding.

However, even the apparently most simple form of legal reasoning – i.e., deductively deciding whether the law covers the specific fact situation – is often more complicated in practice. When looking for the premises of the judicial syllogism, judges are faced with two different sets of problems: examining what happened through evidential reasoning and deciding what consequences flow from what happened through the legal qualification of the fact, both of which are open to the influence of extra-legal factors. In fact, one of the objections that have historically been raised against the syllogistic model is that judges do not decide according to it but do so under the influence of other dynamics like their emotions, idiosyncrasies and preferences of various kinds. According to this view, the syllogistic reasoning would therefore only be an *ex post* rationalisation of a decision already made on other bases. Recognising that judicial decisions are shaped by considerations that go beyond the formal law was a hallmark of Legal Realism – a movement that was concerned primarily with the human factor in legal decision-making.

The basic realist position about judicial decision-making can be said to be a two-part hypothesis (Schauer 2009). The first part consists in the claim that judges have a preferred outcome that precedes consultation of the formal law – the judicial “hunch” (Hutcheson 1929). This preferred outcome can be based on characteristics of the litigant or of the judge, different conceptions of justice, ideology, or assessments of wise policy (legal realists highly diverged on this point), but they all precede the search for a legal justification. The second part consists in the claim that, in looking for the legal justification for their preferred outcome, judges will often find defensible legal justifications available for a wide range of possible outcomes. Realists claimed, then, that in making their decisions, judges respond first to the stimulus of the facts of the case, rather than to legal rules and reasons, and

thus to obtain the best explanation for why judges decide as they do, one ought to look elsewhere (Leiter 2005).

The fact that we now often take the question of what the main determinants of judicial decisions are to be an empirical one is perhaps the most important legacy of the Realist program. For example, insofar as the Attitudinal model explains much of judicial decision-making in terms of judges' political preferences – or attitudes, hence the name attached to the movement –, it can be properly understood as carrying on a realist approach to studying judicial decision-making. Through the use of sophisticated multiple regression techniques, political scientists have been able to ascertain the leading determinants of US Supreme Court outcomes. After analysing a variety of criteria, it has been found that ideology, rather than personality traits, or legal determinations from text and precedent, is the best predictor of that Court's decisions (see, e.g., Landes and Posner 2009; Segal and Spaeth 2004). Similar results have been found in some EU countries' higher courts (Franck 2009; Amaral-Garcia et al. 2009; Garoupa et al. 2013, 2015).

The basic claims of Legal Realism have both contemporary adherents and continuing relevance. As highlighted by Spamann and Klöhn (2016, 3), “realism matters because the specificities of judges and the legal process might well have deep effects on legal reasoning as practiced in courts”. Legal Realist scholars thought of judicial decisions as being predictable but claimed that the key to this prediction lies neither in the consultation of formal legal authorities nor in the internal understanding or self-reports of judges themselves. The prediction is rather best accomplished by determining what factors systematically influence case decisions through empirical (and external) inquiry.

The empirical work on judicial behaviour that emerged in the last decades can be best understood as a new generation of Legal Realism, with researchers conducting what Llewellyn (1931, 1244) and his peers only envisioned: “large-scale quantitative studies of facts and outcome” that assess the influence of the extra-legal factors on case outcomes. In the words of

behavioural law and economic scholars Thomas Miles and Cass Sunstein (2008), the numerous relevant results have produced a “New Legal Realism”, conceptualised as an effort to understand the sources of judicial decisions on the basis of testable hypotheses and large data sets. The distinguishing characteristic of this new strand lies in

the close examination of reported cases in order to understand how judicial ‘personality’, understood in testable ways, influences legal outcomes, and how legal institutions constrain or unleash these influences. These inquiries represent an effort to test certain intuitive ideas about the indeterminacy of law, and to implement the (old-style) realist call for empirical study of how different judges decide cases by responding to the ‘stimulus’ of each case (Miles and Sunstein 2008, 834-5).

In this context, and in light of “a mountain of evidence demonstrating the deficiencies of human reasoning, and little or no evidence that reasoning can perform in the way that rationalist theories (...) require it to perform” (Haidt 2013, 867), some scholars advocate for a new type of Legal Realism, namely “Behavioural Realism” – a school of thought that asks the law to take into account more accurate models of human cognition and behaviour (Kang 2023). In order to do so, it is suggested that legal operators should be on the lookout for a more accurate, upgraded model of human behaviour that comes from the sciences, including experimental social and cognitive psychology, and compare it to the “commonsense” legacy understanding that is embedded within the status quo. When the gap between the two models grows too large, it should be closed by revising the law to take into account the upgraded model. If that cannot be done, either for principled or pragmatic reasons, then lawmakers should articulate the reasons why transparently. Unlike previous theories that have focused on the influence of aspects such as ideology or personal preferences on judges’ decisions, this behaviourist strand of realism examines the influence of factors operating under the level of consciousness,

such as implicit biases, and in doing so also aims to understand the judicial decision-making through the lens of cognitive psychology and neuroscience.

For instance, scholars have suggested to study judicial choice also through the lens of neuroscience, more specifically through the examination of the cognitive processes people use when deciding. This can help illuminate some aspects of decision-making processes that remain hidden, that is, that occur outside of a person's conscious awareness or control, like implicit biases. These hidden processes are part of what is referred to as implicit cognition, which also includes, for example, emotions and empathy, and can influence decisions in ways that we are not aware of. Behavioural, social and neuroscientific findings regarding human decision-making processes not only support the claim that implicit biases are real and present in society but, being applicable to most people, can also inform our understanding about how judges decide (Bradley 2018). In line with this, the next section turns to debiasing strategies designed to address such effects, with a focus on those tested in legal contexts and their respective strengths and shortcomings in counteracting implicit bias in adjudication.

5. Debiasing Strategies in Legal Contexts

Implicit biases, we have seen, are a source of judgment errors. The concept of error generally presupposes a normative standard that dictates how agents should act. As such, biases can be thought of as causes of suboptimal reasoning/decision-making relative to that normative standard. If actual behaviour falls systematically short of normative ideals, the question of how to close this gap naturally emerges. Debiasing measures aim to align actual reasoning/decision-making processes and outcomes with the normative standard, seeking to address the negative effects of biases by improving either the decision-making process or some relevant characteristics of the decision-maker (Zenker 2021). While Section 3 examined the limits of adjudicative debiasing through law – strategies aimed at eliminating or mitigating biases

within adjudication itself, primarily through procedural rules governing how judges or juries ought to decide – this section turns to broader motivational, cognitive, and technological approaches to debiasing.

In his research on judgment and decision-making, psychologist Hal R. Arkes (1998, 449) describes debiasing measures as “all strategies designed to reduce the magnitude of judgment errors”, including errors that may not be owed to biases but to non-conducive environmental, institutional, or cultural conditions, as well as personal constraints such as fatigue and stress. Following Richard Larrick’s (2004) taxonomy, debiasing measures can be analytically classified into motivational, cognitive, or technological. Motivational strategies focus on increasing the individual’s motivation to perform well, for instance through incentives and accountability. They are based on the critical assumption that people possess normative strategies and will use them when the benefits exceed the costs. The two remaining categories, however, do not presume this, but rather assume that, although intuitive strategies are imperfect, they can be replaced by strategies that approach normative standards, even if falling short. Therefore, cognitive techniques focus on modifying the cognitive strategies of individuals through, e.g., training or incentivising them to consider alternative views, and represent a “compromise between a strategy that approximates the normative ideal, but that can be remembered and implemented given ordinary cognitive limitations on memory and computation” (Larrick 2004, 317). Technological debiasing expands “possible strategies to include techniques external to the decision maker”, recognising that “individual reasoning can approach normative standards through the use of tools” (Larrick 2004, 318). Examples include improving information processing through decision aids and even replacing individual judgment with statistical models or artificial intelligence.

Researchers have explored the effectiveness of various debiasing techniques, including in legal contexts. These techniques encompass individual-level strategies that judges can adopt themselves to mitigate implicit biases, and institutional-level reforms that could be more widely

implemented in the judiciary, also referred to as debiasing infrastructures (Zenker 2021, 397). In other words, they target both individual dispositions and abilities, as well as the structure of collective environments. Following Larrick's taxonomy, these debiasing strategies seek to offer subjects personal incentives to use a more desirable mode of reasoning and decision-making, where doing so would be apt; to provide environmental props that facilitate the deployment of these processes as a matter of standardised procedures; and to personally provide them with the knowledge, skill, or information necessary to properly deploy such reasoning processes.

The first step in overcoming implicit biases is to bring them to the awareness of the decision-maker. While many people nowadays recognise the existence and impact of most of the biases described by social and cognitive psychologists, they are less aware of the impact that they have in governing their own judgments (due in part to the bias blind spot). Creating general bias awareness is a debiasing technique that involves educating decision-makers about the existence of implicit biases and how they can affect the decision-making process. This basic education process, which has been applied in a variety of fields, generally takes the form of brief, nontechnical, and intensive tutorials in which specific biases are demonstrated to decision-makers in general terms. Through a better understanding of the underlying decision mechanisms, this strategy seeks to lessen the susceptibility of decision-makers to biases (Reese 2012, 1281).

Courses and training sessions targeting implicit bias have gained momentum in the last years within judicial institutions. For example, some American states such as New York and California have required that sitting judges and practicing lawyers include credit hours of diversity and inclusion aimed at eliminating biases as part of their continuing legal education (Breger 2019). The American Bar Association (2016) issued a Resolution encouraging all courts that currently require mandatory continuing legal education to modify their rules to include, as a separate required credit, programs regarding diversity and inclusion in the legal profession, and

programs regarding the elimination of bias. Likewise, the Brennan Centre for Justice recommends implicit bias training for judges, as well as training for those who are tasked with selecting judges (Berry 2016). Rachlinski and colleagues (2009, 1228) suggest that this strategy could be furthered, for instance, by the use of judges' Implicit Association Test (IAT) scores, which might "help newly elected or appointed judges understand the extent to which they have implicit biases and alert them to the need to correct for those biases on the job", and therefore enable the system to provide targeted training about bias to new judges. The same could be done with Cognitive Reflection Test (CRT) scores, which could be used to raise awareness to cognitive biases. While these scores may not be able to forecast individual behaviour, they can surely make explicit to judges that they are also vulnerable to bias.

The problem of bias awareness courses and seminars is that they often come unaccompanied with other debiasing interventions. In other words, decision-makers are made aware of their cognitive shortcomings but are given no information or guidance on what to do next. Here, an additional debiasing strategy that could be implemented in the judiciary is auditing. Auditing judges' decisions could not only help them understand the extent to which their work is being influenced by implicit biases (thus raising awareness), but it also works as a motivational strategy, since it improves judicial accountability. Motivational debiasing strategies focus on increasing the individual's motivation to perform well, for instance through incentives and accountability.

As individual judges exercise their daily discretion, it will often be impossible to determine in any specific situation whether an implicit bias played a causal role. If, however, similar decisions are recorded over time and/or by multiple decision makers, the data may reveal patterns of bias, helping identify areas of concern that warrant deeper examination. There is already a great deal of aggregate data regarding the ways in which the race, gender, and other demographic features of litigants affect, e.g., pretrial detention, sentencing, summary judgment motions, and other related issues.

Aggregate data regarding the influence of judges' demographic traits on their rulings is also available. There is, however, a lack of "data about what individual judges are doing that might enable them to better calibrate their decisions. This makes it easier for individual judges to deny that they are part of the problem" (Wistrich and Rachlinski 2017, 108). The justice system would benefit from the implementation of auditing programs to evaluate the decisions of individual judges, especially regarding discretionary matters such as bail-setting, sentencing, or child-custody allocation, in order to determine whether they appear to be influenced by implicit bias. Jerry Kang (2021) also suggests auditing in his research on implicit biases, saying that judges could, at least at the individual level, collect data to assess their patterns in exercising judicial power, especially regarding decisions involving substantial discretion. Auditing can help mitigate the problem by identifying the areas where intervention might be necessary. Besides, it can help to overcome the first challenge to debiasing, namely, that of the bias blind spot and people's lack of awareness regarding their own biases.

Cognitive debiasing techniques, in their turn, focus on modifying the cognitive strategies of individuals through training and the adoption of better decision-making strategies. Training opportunities for judges going beyond their continuing legal education could be offered, including training in rules and representations. One strategy for improving decision-making is to teach people the necessary rules and principles required to do so. Inferior reasoning strategies can, with training, be replaced by better strategies. Based on Richard Nisbett's (1993) extensive research program exploring the effectiveness of training on normative rules, Larrick (2004, 318) identifies two sets of implications for debiasing: first, there are specific cognitive factors that facilitate the learning and use of normative rules; and second, formal training in basic disciplines, such as economics and statistics, is an important cultural mechanism for transmitting effective cognitive strategies. Judges can also be taught through training how to replace inferior reasoning strategies with better ones, learning, for instance, how to correctly adopt

strategies such as consider-the-opposite and perspective-taking. The former strategy requires a person to imagine and explain the basis for alternate outcomes, especially those that conflict with the opinion that he or she holds (Wistrich and Rachlinski 2017). This technique has been shown to be effective at combating cognitive biases such as hindsight, anchoring, overconfidence and confirmation bias (see, e.g., Adame 2016; Van Brussel et al. 2020; Sanna and Schwarz 2003; Mussweiler et al. 2000). A somewhat similar strategy is “the devil’s advocate”, which formalises the dissent process by bringing in a second person to question the decision maker’s conclusion (Reese 2012).

Diversity training can provide judges with the opportunity to challenge their assumptions about different social groups, exposing them to stereotypical-incongruent models. Exposing individuals to examples and narratives that challenge existing stereotypes can help to break down stereotypical associations and promote more accurate perceptions. Encountering a group member whose characteristics contradict the correspondent group stereotype can influence perceptions of the group, thus constituting a valuable learning experience (Garcia-Marques and Mackie 1999). Among interventions designed to reduce social biases, exposure to counter stereotypical exemplars has been repeatedly found to be one of the most promising techniques (FitzGerald et al. 2019; Forscher et al. 2019; Lai et al. 2014). Nonetheless, consciously attempting to change implicit associations in judicial settings might be difficult, as Wistrich and Rachlinski (2017) explain, given that most judges have little control over their dockets, which tend to include an over-representation of black criminal defendants. This frequent exposure contributes to perpetuating negative associations with black individuals. To avoid this scenario, the authors suggest that courts should consider rotating judges among specialist assignments so that implicit negative attitudes formed while deciding criminal cases will not take root.

Finally, there are also technological debiasing strategies that could be adopted in judicial settings, which encompass techniques external to the

decision-maker, providing environmental props that facilitate the adoption of standardised procedures. Checklists, for instance, can help cabin discretion in ways that increase overall accuracy (Ashton 1992). Checklists are a potent tool for streamlining processes and thus reducing errors, especially those arising from forgetfulness and other memory distortions, e.g. over-reliance on the availability heuristic (Soll, Milkman, and Payne 2015). The adoption of checklists qualifies as a technological debiasing measure by offering agents with decision-making props, having primarily served areas such as aviation, healthcare, and product manufacturing, where best practices are likely to be overlooked due to complexity, time pressures, high stress, or fatigue. They have “a long history of improving decisions in high-stakes contexts and are particularly well suited to preventing the repetition of past errors” (Kahneman et al. 2021, 218). The adoption of checklists has been suggested as a way to free judges from reliance on their memories and encourage them to proceed methodically, thereby ensuring that they touch all of the deliberative bases. By reviewing a checklist at each step in the decision-making process, a judge is less likely to rely on intuition when doing so is inadvisable (Guthrie, Rachlinski, and Wistrich 2007, 138). Besides, individuals attend to and process information more comprehensively when they have a mental schema that tells them what information is needed in a given situation and where to find it (Heath et al. 1998, 15). Constraining decision-making by adopting checklists or similar tools like scripts thus tend to lead to more accurate and consistent decisions – a recommendation in line, for example, with the structured interview literature (see Kang et al. 2012).

Overall, the main debiasing strategies available to the judiciary include motivational approaches, such as accountability mechanisms and auditing; cognitive techniques, including training, perspective-taking, and exposure to counter-stereotypical exemplars; and technological supports, like checklists and structured procedures. Each of these interventions offers distinct advantages and limitations, but collectively they illustrate the range of possibilities for reducing the influence of implicit bias on judicial reasoning.

In line with this body of research, and consistent with the aims of the behavioural realist movement, the following section will set out recommendations for mitigating implicit biases in judicial decision-making.

6. Improving Judicial Capabilities

In keeping with the objectives of the Behavioural realist movement, and in line with the existing body of research on debiasing techniques, a number of general recommendations aiming at mitigating judicial implicit biases can be made. The first recommendation is to audit judicial decisions, especially those involving substantial discretion. As individual judges exercise their daily discretion, it will often be impossible to determine in each and any situation whether implicit bias played a causal role. If, however, similar decisions are recorded over time and/or by multiple decision makers, the data may reveal patterns of bias, helping identify areas of concern that warrant deeper examination. Not only can auditing pinpoint the areas where intervention might be necessary, it can also raise judicial awareness to their own biases, enabling judges to overcome overconfidence and the bias blind spot. In addition, auditing can provide them with valuable feedback about their decision-making, which they currently lack, since the institutional context in which judges operate, which can be classified as a wicked learning environment, lacks appropriate feedback about their decisions. Finally, auditing also incentivises accountability, and hence serves as a motivational strategy, since the prospect of having a board of peers assessing one's decisions can encourage cognitively complex reasoning and self-awareness.

This process could be conducted by an audit panel, composed mainly by a diverse group of judges and court staff, who would be responsible for evaluating the consistency and objectivity within a pool of judicial decisions, with the goal of detecting disparities which may have been occasioned by implicit bias. In order to ensure that this auditing process is not itself affected by the evaluators' implicit bias, a blind review format, in which a sample of

anonymised decisions are assigned to members of the panel, should be adopted. This step could be facilitated by the use of artificial intelligence, which outperforms human evaluators in identifying hidden patterns of implicit bias in a large number of judicial decisions – and here, the growing digitalisation of court proceedings would be helpful in providing input for the system. However, further research on how, precisely, to build this model is still necessary. For instance, issues related to classifying a decision as biased, and overall concerns about the opacity of some AI systems, would need to be addressed (for a discussion, see Lopes 2024). If such a system were to be built and employed to detect patterns of bias in judicial decisions, it would be important that it be overseen by the audit panel. The panel would be responsible for verifying its accuracy and analysing the causes for the disparities detected. Only then could they ascertain whether corrective interventions should be adopted. Importantly, auditing should be conducted for educational purposes and for gaining insight into how implicit biases are affecting judicial decisions, helping target improvement strategies such as training. It should in no way be used to justify disciplinary proceedings, or to penalise judges whatsoever.

The second recommendation is to regularly screen judges, especially those above a certain age, for cognitive decline. Research measuring the natural decline of cognitive performance due to ageing generally indicates that, starting from the age of 65, key cognitive abilities involved in areas such as judgment, decision-making, problem solving, attention, memory, and planning – all of which are important skills for the effective discharge of the judicial function – naturally start to deteriorate. Considering how, in most jurisdictions, judges are appointed for life, it is important to ensure that their cognitive capabilities are up to par, and they are fit for duty until the legal retirement age is reached. Judicial screening should be conducted by physicians specialising in areas such as neurology and psychiatry, who would examine judges for any type of cognitive impairment which could jeopardize their performance. While it is likely that most judges above a certain age

already undergo some form of regular checkup by their personal physicians, these are bound by rules of confidentiality and may lack the expertise necessary to detect intellectual decline. These evaluations would thus be best conducted by independent experts, who are hired by and report to a judicial oversight board. Similar screenings are done by aviation medical examiners, who conduct physical examinations to assess the “fitness of flight” of pilots, and are trained and designated by an administrative aviation organ (see Matthews and Stretanski 2024).

The screening of judges for cognitive decline would have to take place within an appropriate framework and infrastructure, to be developed by the oversight board to ensure that appropriate safeguards are in place to guarantee a fair procedure, for instance, by providing that judges who fail an examination have the right to appeal. Moreover, it would have to establish how often these assessments would take place, and if they would target only elderly judges or also – albeit perhaps less frequently – younger or middle-aged judges. Additionally, failing an examination should not result in an immediate assumption of unfitness for duty, and the board should verify, based on information concerning the judge’s professional performance, e.g., courtroom behaviour, rate of overturned decisions, whether or not additional measures should be taken.

On a similar note, it is recommended that judges undergo regular health evaluations to assess their physical and mental wellbeing, which are intimately connected to the proper performance of their duties. As previously observed, judges, like most actors in the judicial system, are stressed, overburdened, and operating under time pressures. The most common contributing factor to judicial stress seems to be excessive workloads, and commonly reported negative consequences associated therewith are judgment errors, as well as diminished cognitive abilities, lack of concentration, reduced reasoning skills and clarity of thought. Empirical evidence suggests that stress leads people to scan alternatives less systematically and completely, and that time pressures are correlated with less accurate

decisions. Furthermore, stress leads to fatigue, low energy and sleep disturbances, all of which can have a detrimental effect on complex cognitive functions. Yearly or biannual evaluations by physicians and psychologists can therefore help determine the state of judges' wellbeing and assess possible interventions that can ameliorate the problem (e.g., professional counselling or therapy, stress-management and emotional resilience courses, wellness leaves). During these evaluations, judges can also receive personalised guidance regarding dietary and exercise interventions aimed at improving cognitive performance. In addition, regular health evaluations contribute to raising institutional awareness to the negative impact of stress and mental health issues within the judicial function, removing stereotypes and stigmatisation still often associated with this topic. Importantly, an investigation of the values and practices within judiciaries that lead to the promotion of unhealthy efficiency is also necessary. Looking into the root causes of judicial stress and promoting changes in the judicial work culture through a better distribution of caseload among courts, monitoring judges' workload, including controlled rest times for mental and physical exercises, and incentivising small active breaks during working hours can help reduce judgment errors and are worth investing in.

The fourth recommendation concerns the incorporation of debiasing strategies in judges' ongoing education and training curricula. This includes the promotion of bias awareness, bringing into light the influence that implicit bias has on judicial decisions, irrespective of judges' good intentions or commitments to objectivity and fairness. Initiatives aiming at promoting bias awareness are already being implemented in some jurisdictions, and should be a mandatory component of judicial education. There are, however, two challenges related to these kinds of initiatives: first, despite raising awareness to the overall existence and impact of implicit biases in judicial decision-making, judges' may remain unconvinced that they are also personally affected by them. This problem could be addressed by presenting judges with their scores from the Implicit Association Test (IAT) and the Cognitive

Reflection Test (CRT), as well as by auditing their decisions, as mentioned above. The second challenge lies in the fact that bias awareness initiatives often tend not to be accompanied by other interventions. In other words, decision-makers are made aware of their cognitive shortcomings, but are given no information or guidance on what to do next. As previously argued, raising awareness is insufficient if not accompanied by giving decision-makers the tools for overcoming their biases.

Therefore, beyond bias awareness programs, judiciaries should offer specific training sessions focusing on teaching judges reasoning skills that are helpful in combating different kinds of cognitive bias. Examples include training in rules and representations, encompassing formal statistical and logical training, and teaching judges better reasoning strategies like consider-the-opposite. Concerning the latter, besides teaching judges to consider alternative scenarios and encouraging them to do so in their daily professional lives, judiciaries could also implement case review procedures, through peer reviews, panel discussions, or structured deliberation processes (like checklists). When it comes to combating social bias, the judiciary should promote inclusion and diversity training, which exposes judges to stereotype-incongruent models and allows them to assume the perspective of others, increasing empathy and reducing implicit associations and stereotypes towards certain groups. This kind of training can have its effectiveness increased by virtual reality technologies, with tech companies already providing perspective-taking training in VR.

Judicial implicit bias should be approached through a multimodal perspective. Existing debiasing interventions tend to produce limited effects, most likely due to the complex nature of this type of bias and the need to address aspects of cognition, motivation, and technology. Therefore, the best approach is to combine complementary interventions to target each of these components, and that can collectively make up for the limitations of individual strategies. Considering the complexity of implicit bias and of

human cognition, the adoption of multiple techniques in parallel is the most promising.

Finally, these recommendations should be framed as a legal obligation, by explicitly including them in the scope of the judicial professional duties. This is important to ensure that these measures “have teeth”, thus guaranteeing that a failure to comply with them could lead to disciplinary sanctions. Concerning the latter, since the specific violations which may give rise to it ought to be defined in advance in professional codes of conduct, it is necessary that the suggested interventions aimed at improving judicial cognitive functioning and performance be explicitly incorporated into these codes, and not merely implicitly interpreted from other broad judicial duties. This would allow for a failure to accept these interventions without a proper justification for doing so to be interpreted as a grossly negligent behaviour.

Recommendation CM/Rec (2010)12 of the Council of Europe on judges' independence, efficiency, and responsibilities, defines that judges should be guided in their activities by ethical principles of professional conduct, which should be laid down by EU Member-States in codes of judicial conduct. Although these codes are often referred to as codes of ethics, the duties laid down by them have a (quasi-) legal character, referring to a responsibility that a professional has on the basis of a prior regulatory framework. Codes of judicial conduct normally describe the professional obligations of judges, and sometimes also determine a disciplinary regime in view of non-compliance therewith. While professional duties aim at achieving, in an optimal manner, the best professional practices, disciplinary regimes are essentially meant to sanction failures in the accomplishment of (at least some of) these duties. These professional duties can be justified in terms of Hart's concept of role-responsibility (2008, 212): whenever a person occupies a distinctive position within a social organisation, to which distinct duties are attached to provide for the welfare of others or to advance in some specific way the aims or purposes of the organisation, he or she can be said to be responsible for the performance of these duties, or for doing what is necessary to fulfil them.

While specific obligations may vary according to jurisdiction, some common duties are present in virtually all codes seeking to discipline the judicial role. One judicial duty commonly encountered in codes of conduct is the maintenance of judges' overall fitness as good adjudicators, having as an end goal the guarantee of the proper application of the law in an impartial, just, fair and efficient manner (CCJE 2010). In this regard, Recommendation CM/Rec (2010)¹² emphasises, for instance, how “judges should regularly update and develop their proficiency” (2010, 15). In the explanatory memorandum that accompanies the Recommendation, it is explained that judges can do so by attending training programs, but also “through personal efforts to obtain the knowledge and skills required to continuously provide quality justice” (CoE 2010, 32). On a similar note, the ECtHR's Resolution on Judicial Ethics (2021) states that judges should strive to enhance their professional knowledge and skills to maintain a high level of competence.

The idea that judges ought to take action towards improving the knowledge and abilities that are necessary to continuously provide quality justice provides support to judges' responsibility to adopt certain interventions aimed at mitigating implicit biases. There seems to be a general principle establishing that “there is an obligation not just to have a basic level of competence but also to take active steps to enhance the skills and qualities necessary for the proper performance of judicial duties” (Chandler and Dodek 2016, 339). If cognitive limitations are preventing judges from fulfilling the duties that their role requires, such as impartiality, one can argue that they must take the necessary steps to improve their capabilities.

In other words, it is recognised that the judicial role comes with a responsibility to fulfil certain professional duties, including the duty to acquire and enhance the knowledge and skills necessary for carrying out the judicial role in a proper manner. Jennifer Chandler and Adam Dodek (2016) have defended a similar approach, based on common principles enshrined in the United Nations' Office on Drugs and Crime (UNODC 2002) Bangalore Principles of Judicial Conduct, endorsed by the UN Human Rights

Commission in 2003, and recognised by various legal systems worldwide. Among the core duties recognised in the document is that of competence and diligence, according to which judges must take reasonable steps to maintain and enhance the knowledge, skills and personal qualities necessary for the proper performance of judicial duties. For this purpose, they should take advantage of the training and other facilities which should be made available, under judicial control, to judges (UNODC 2002).

What is contemplated by current codes of conduct and judicial practice is reasonable care for one's mental and physical health and the pursuit of certain educational opportunities. This can encompass some strategies aimed at reducing bias: while a duty to undergo regular psychological and medical evaluations, to submit oneself to dietary and exercise interventions and to undergo behavioural cognitive training could be derived from the legal duty of care for one's mental and physical health, a duty to participate in training targeting implicit biases and to adopt checklists and other decision props could be derived from the legal duty to foster one's education.

While the preceding recommendations provide a framework for strengthening judicial capabilities and mitigating implicit bias, their practical feasibility and effectiveness raise some unresolved questions. A first limitation concerns the current state of evidence regarding debiasing interventions. Although various strategies – ranging from bias awareness initiatives to auditing mechanisms and technological aids – have shown promise, the available empirical findings are mixed, with some studies reporting limited effects. Systematic evaluation of debiasing methods, including critical consideration of negative findings, remains necessary to assess whether these interventions can achieve sustained improvements in judicial decision-making. A related issue is the possibility of empirically verifying the effectiveness of proposed solutions within judicial contexts. Most evidence to date has been gathered in experimental or non-legal environments, raising concerns about the extent to which these results can be generalised to courts. Robust methodologies, pilot projects, and case studies

in actual judicial settings would be required to evaluate both their suitability and their limits. Closely related is the question of context-specificity: some interventions may prove useful only under particular institutional or cultural conditions, and therefore their transferability across jurisdictions cannot be assumed. Moreover, attention must be given to potential negative side effects and practical barriers to implementation. Interventions such as mandatory screenings, regular health checks, or extensive training programs may raise concerns about costs, administrative feasibility, the limited time available to judges, and possible resistance within the judiciary. Even if formally adopted, their long-term effectiveness would need to be monitored and evaluated in practice.

These considerations highlight the need for more sustained empirical and normative inquiry into the design, testing, and implementation of debiasing interventions in judicial contexts, particularly before any of the proposed strategies are implemented, and especially prior to framing them as duties incumbent upon judges. A comprehensive analysis of these questions falls outside the scope of this article, but they remain important avenues for further investigation in future scholarly work.

7. Conclusions

This article has explored the gap between the normative expectations of judicial decision-making and the empirical reality of how judges actually decide cases. While the legal system operates under the assumption that judges follow a rational, syllogistic model of reasoning – moving linearly from facts to law to verdict –, decades of research in cognitive and social psychology reveal a more complex picture. Judges, despite their training and commitment to impartiality, remain susceptible to the pervasive influence of implicit biases that operate below the level of conscious awareness. The analysis presented here demonstrates that traditional legal safeguards, though valuable, are fundamentally insufficient to address implicit bias because they

were designed to combat explicit instances of prejudice and partiality. Current procedural protections, including the right to appeal, requirements for reasoned verdicts, and rules governing judicial recusal, assume a rationalist model of judicial reasoning that fails to account for implicit biases.

Drawing on the insights of Behavioural Realism, this article has argued that the legal system must evolve to incorporate more accurate models of human cognition and behaviour. The evidence presented reveals how various cognitive and social biases consistently influence judicial outcomes across different jurisdictions and throughout different stages of legal proceedings. Moreover, environmental factors such as judicial fatigue, time pressure, and the natural decline of cognitive abilities with age further compound these biasing effects, creating conditions particularly conducive to systematic decision-making errors.

To address these challenges, this article has proposed a comprehensive framework of debiasing interventions that operate at both an individual and at an institutional level. These strategies range from motivational approaches (such as auditing judicial decisions to enhance accountability) to cognitive techniques (including bias awareness training and teaching judges to consider alternative scenarios) to technological solutions (such as the implementation of decision-support checklists). The multimodal approach advocated here recognises that no single intervention is sufficient to address the complex nature of implicit bias, requiring instead a coordinated effort that targets cognitive, motivational, and technological dimensions simultaneously.

These debiasing interventions should not remain optional enhancements to judicial practice but should be encompassed within judges' existing professional duty of competence and diligence. By explicitly incorporating these measures into judicial codes of conduct, the legal system can ensure that efforts to mitigate implicit bias have the necessary institutional support and enforcement mechanisms. This approach builds upon established principles already recognised in judicial ethics codes worldwide, extending the duty to maintain professional competence to include the adoption of evidence-based

strategies for improving decision-making capabilities. At the same time, however, further research is needed to evaluate the practical feasibility, effectiveness, and limitations of these measures before they can justifiably be demanded as binding professional duties to judges.

The recommendations presented here constitute a preliminary set of measures for judicial systems to address cognitive and social biases, serving as a foundation for future research and more comprehensive reform. They acknowledge that, beyond legal expertise, an understanding of and response to the inherent limitations of decision-makers cognitive capabilities is also necessary. Given the profound consequences of judicial decisions for individuals and society, ensuring the accuracy and fairness of these decisions represents a fundamental duty of the justice system. The failure to address implicit biases not only undermines individual cases but also erodes public confidence in the judicial system's commitment to equal treatment under the law.

References

- Adame B. (2016). Training in the Mitigation of Anchoring Bias: A Test of the Consider-the-Opposite Strategy, in *Learning and Motivation*, n. 53.
- Amaral-Garcia S., Garoupa S. and Grembi V. (2009). Judicial Independence and Party Politics in the Kelsenian Constitutional Courts: The Case of Portugal, in *Journal of Empirical Legal Studies*, n. 6.
- Arkes H. R. (1998). Principles in Judgment/Decision-Making Research Pertinent to Legal Proceedings, in *Behavioral Sciences & the Law*, n. 7.
- Arnold D., Dobbie W. and Hull P. (2020). Measuring Racial Discrimination in Bail Decisions, in *NBER Working Paper Series*. https://www.nber.org/system/files/working_papers/-w26999/w26999.pdf.
- Ashton R. H. (1992). Effective Justification and a Mechanical Aid on Judgment Performance, in *Organizational Behavior and Human Decision Processes*, n. 52.

- Axt J., Nguyen H. and Nosek B. (2018). The Judgment Bias Task: A Flexible Method for Assessing Individual Differences in Social Judgment Biases, in *Journal of Experimental Social Psychology*, n. 76.
- Berry K. (2016). Building a Diverse Branch: A Guide for Judicial Nominating Commissioners, in *Brennan Center for Justice*,
https://www.brennancenter.org/sites/default/files/publications/Building_Diverse_Bench.pdf.
- Bielen S., Marneffe W. and Mocan N. (2021). Racial Bias and In-Group Bias in Virtual Reality Courtrooms, in *The Journal of Law & Economics*, n. 64.
- Bradley A. (2018). The Disruptive Neuroscience of Judicial Choice, in *UC Irvine Law Review*, n. 9.
- Braun R. (2000). The Effect of Time Pressure on Auditor Attention to Qualitative Aspects of Misstatements Indicative of Potential Fraudulent Financial Reporting, in *Accounting, Organizations and Society*, n. 25.
- Breger M. (2019). Making the Invisible Visible: Exploring Implicit Bias, Judicial Diversity, and The Bench Trial, in *University of Richmond Law Review*, n. 53.
- Bystranowski P., Janik B., Próchnicki M. and Skórska P. (2021). Anchoring Effect in Legal Decision-Making: A Meta-Analysis, in *Law and Human Behavior*, n. 45.
- Canale D. and Tuzet G. (2020). *La giustificazione della decisione giudiziale* (G. Giappichelli Editore).
- Chandler J. and Dodek A. (2016). Cognitive Enhancement in the Courtroom, in F. Jotterand and V. Dubljevic (eds.) *Cognitive Enhancement* (Oxford University Press).
- Cho K., Barnes C. and Guanara C. (2017). Sleepy Punishers Are Harsh Punishers: Daylight Saving Time and Legal Sentences, in *Psychological Science*, n. 28.
- Cohen M. (2015). When Judges Have Reasons Not to Give Reasons: A Comparative Law Approach, in *Washington & Lee Law Review*, n. 72.

- Cohen R., Lind A. and Tyler T. (1989). The Social Psychology of Procedural Justice, in *Contemporary Sociology*, n. 18.
- Consultative Council for European Judges (CCJE) (2010). *Magna Carta of Judges* (Council of Europe).
- Council of Europe (1984). *Protocol No. 7 to the Convention for the Protection of Human Rights and Fundamental Freedoms, as Amended by Protocol No. 11* (Council of Europe).
- Council of Europe (2010). *Judges: Independence, efficiency and responsibilities*, Recommendation CM/Rec(2010)12 (Council of Europe).
- Court of Justice of the European Union (CJEU) (2016). *Code of Conduct for Members and former Members of the Court of Justice of the European Union* (2016/C 483/01). (Court of Justice of the European Union).
- Croskerry P., Singhal G. and Mamede S. (2013). Cognitive Debiasing 1: Origins of Bias and Theory of Debiasing, in *BMJ Quality & Safety*, n. 22.
- Danziger S., Levav J. and Avnaim-Pesso L. (2011). Extraneous Factors in Judicial Decisions, in *Proceedings of the National Academy of Sciences*, n. 108.
- Eisenberg M. (2022). *Legal Reasoning* (Cambridge University Press).
- Englich B., Mussweiler T. and Strack F. (2006). Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making, in *Personality and Social Psychology Bulletin*, n. 32.
- Enough B. and Mussweiler T. (2001). Sentencing Under Uncertainty: Anchoring Effects in the Courtroom, in *Journal of Applied Social Psychology*, n. 31.
- Fitzgerald C., Martin A., Berner D. and Hurst S. (2019). Interventions Designed to Reduce Implicit Prejudices and Implicit Stereotypes in Real World Contexts: A Systematic Review, in *BMC Psychology*, n. 7.
- Flanagan D., Ortiz S. and Alfonso V. (2013). *Essentials of Cross-Battery Assessment* (Wiley).

- Forscher P., Lai C., Axt J. et al. (2019). A Meta-Analysis of Procedures to Change Implicit Measures, in *Journal of Personality and Social Psychology*, n. 117.
- Franck R. (2008). Judicial Independence Under a Divided Polity: A Study of the Rulings of the French Constitutional Court, 1959-2006, in *Journal of Law, Economics, and Organization*, n. 25.
- Garcia-Marques L. and Mackie D. (1999). The Impact of Stereotype-Incongruent Information on Perceived Group Variability and Stereotype Change, in *Journal of Personality and Social Psychology*, n. 77.
- Garoupa N. and Grembi V. (2012). Judicial Review and Political Bias: Moving from Consensual to Majoritarian Democracy, available at *SSRN Electronic Journal*: <https://ssrn.com/abstract=2097259>.
- Garoupa, N., Gomez-Pomar F. and Grembi V. (2013). Judging under Political Pressure: An Empirical Analysis of Constitutional Review Voting in the Spanish Constitutional Court, in *Journal of Law, Economics, & Organization*, n. 29.
- Guthrie C., Rachlinski J. and Wistrich A. (2009). The 'Hidden Judiciary': An Empirical Examination of Executive Branch Justice, in *Duke Law Journal*, n. 58.
- Guthrie C., Rachlinski J. and Wistrich A. (2007). Blinking on the Bench: How Judges Decide Cases, in *Cornell Law Review*, n. 93.
- Guthrie C., Rachlinski J. and Wistrich A. (2001). Inside the Judicial Mind, in *Cornell Law Review*, n. 86.
- Haidt J. (2013). Moral Psychology and the Law: How Intuitions Drive Reasoning, Judgment, and the Search for Evidence, in *University of Alabama Law Review*, n. 64.
- Hart H. L. A. (2008). *Punishment and Responsibility: Essays in the Philosophy of Law* (Oxford University Press).
- Heath C., Larrick P. and Klayman J. (1998). Cognitive Repairs: How Organizational Practices Can Compensate for Individual Shortcomings, in *Research in Organizational Behaviour*, n. 20.

- Hogarth R. (2010). *Educating Intuition* (University of Chicago Press).
- Hogarth R., Lejarraga T. and Soyer E. (2015). The Two Settings of Kind and Wicked Learning Environments, in *Current Directions in Psychological Science*, n. 24.
- Hutcheson J. (1929). The Judgment Intuitive: The Function of the 'Hunch' in Judicial Decision, in *Cornell Law Journal*, n. 14.
- Jolls C. and Sunstein C. (2006). Debiasing through Law, in *The Journal of Legal Studies*, n. 35.
- Kahneman D., Sibony O. and Sunstein C. (2021). *Noise: A Flaw in Human Judgment* (Little, Brown Spark).
- Kang J. (2023). Judicial Behavioural Realism about Implicit Bias, in R. Hollander-Blumoff (ed.), *Research Handbook in Psychology and Law* (Edward Elgar).
- Kang J. (2021). What Judges Can Do about Implicit Bias, in *Court Review*, n. 57.
- Kaufman A. (2021). The Precipitous Decline in Reasoning and Other Key Abilities with Age and Its Implications for Federal Judges, in *Journal of Intelligence*, n. 9.
- Kaufman A., Salthouse T., Scheiber C. and Chen H. (2016). Age differences and educational attainment across the lifespan on three generations of Wechsler adult scales, in *Journal of Psychoeducational Assessment*, n. 34.
- Keinan G. (1987). Decision Making under Stress: Scanning of Alternatives under Controllable and Uncontrollable Threats, in *Journal of Personality and Social Psychology*, n. 52.
- Lai C., Marini M., Lehr S. et al. (2014). Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions, in *Journal of Experimental Psychology General*, n. 143.
- Landes W. and Posner R. (2009). Rational Judicial Behavior: A Statistical Study, in *Journal of Legal Analysis*, n. 1.
- Larrick R. (2004). Debiasing, in D. Koehler and N. Harvey (eds.), *Blackwell Handbook of Judgment and Decision Making* (Blackwell).

- Leibovitch A. (2021). Institutional Design and the Psychology of the Trial Judge, in B. Brožek, J. Hage and N. Vincent (eds.), *Law and Mind* (Cambridge University Press).
- Leiter B. (2005). American Legal Realism, in M. Golding and W. Edmundson (eds.), *The Blackwell Guide to the Philosophy of Law and Legal Theory* (Wiley).
- Lidén M., Gräns M. and Juslin P. (2019). 'Guilty, No Doubt': Detention Provoking Confirmation Bias in Judges' Guilt Assessments and Debiasing Techniques, in *Psychology, Crime & Law*, n. 25.
- Liu Z. (2018). Does Reason Writing Reduce Decision Bias? Experimental Evidence from Judges in China, in *The Journal of Legal Studies*, n. 47.
- Llewellyn K. (1931). Some Realism about Realism: Responding to Dean Pound, in *Harvard Law Review*, n. 44.
- Maroney T., Swenson D., Bibelhausen J. and Marc D. (2023). The State of Judges' Well-Being: A Report on the 2019 National Judicial Stress and Resiliency Survey, in *Bolch Judicial Institute at Duke Law*, n. 107.
- Matthews M. and Stretanski M. (2024). *Pilot Medical Certification* (StatPearls Publishing).
- Miles T. and Sunstein C. (2008). The New Legal Realism, in *University of Chicago Law Review*, n. 75.
- Miller A. (2019). Expertise Fails to Attenuate Gendered Biases in Judicial Decision-Making, in *Social Psychological and Personality Science*, n. 10.
- Mitchell G. (2002). Why Law and Economics' Perfect Rationality Should Not Be Traded for Behavioural Law and Economics' Equal Incompetence, available at *SSRN Electronic Journal*: <https://ssrn.com/abstract=306562>.
- Mussweiler T., Strack F. and Pfeiffer T. (2000). Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility, in *Personality and Social Psychology Bulletin*, n. 26.
- Nisbett R. (1993). *Rules for Reasoning* (Lawrence Erlbaum Associates).

- Oeberst A. and Goeckenjan I. (2016). When Being Wise after the Event Results in Injustice: Evidence for Hindsight Bias in Judges' Negligence Assessments, in *Psychology, Public Policy, and Law*, n. 22.
- Pignatiello G., Martin R. and Hickman R. (2020). Decision Fatigue: A Conceptual Analysis, in *Journal of Health Psychology*, n. 25.
- Rachlinski J. and Wistrich A. (2021). Benevolent Sexism in Judges, in *San Diego Law Review*, n. 58.
- Rachlinski, J., Wistrich A. and Guthrie C. (2015). Can Judges Make Reliable Numeric Judgments? Distorted Damages and Skewed Sentences, in *Indiana Law Journal*, n. 90.
- Rachlinski J., Johnson S., Wistrich A. and Guthrie C. (2009). Does Unconscious Racial Bias Affect Trial Judges?, in *Notre Dame Law Review*, n. 84.
- Rassin E. (2020). Context Effect and Confirmation Bias in Criminal Fact Finding, in *Legal and Criminological Psychology*, n. 25.
- Reese E. (2012). Techniques for Mitigating Cognitive Biases in Fingerprint Identification, in *UCLA Law Review*, n. 59.
- Sanna L. and Schwarz N. (2003). Debiasing the Hindsight Bias: The Role of Accessibility Experiences and (Mis)Attributions, in *Journal of Experimental Social Psychology*, n. 39.
- Schauer F. (2009). *Thinking like a Lawyer: A New Introduction to Legal Reasoning* (Harvard University Press).
- Segal J. and Spaeth H. (2002). *The Supreme Court and the Attitudinal Model Revisited* (Cambridge University Press).
- Shroff R. and Vamvourellis K. (2022). Pretrial Release Judgments and Decision Fatigue, in *Judgment and Decision Making*, n. 17.
- Soll J., Milkman K. and Payne J. (2015). A User's Guide to Debiasing, in G. Keren and G. Wu (eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (Wiley & Sons).

Spamann H. and Klöhn L. (2016). Justice Is Less Blind, and Less Legalistic, than We Thought: Evidence from an Experiment with Real Judges, in *The Journal of Legal Studies*, n. 45.

Thaman S. (2019). Appeal and Cassation in Continental European Criminal Justice Systems: Guarantees of Factual Accuracy, or Vehicles for Administrative Control?, in D. Brown, J. Turner and B. Weisser (eds.), *The Oxford Handbook of Criminal Process* (Oxford University Press).

Torres L. and Williams J. (2022). Tired Judges? An Examination of the Effect of Decision Fatigue in Bail Proceedings, in *Criminal Justice and Behavior*, n. 49.

United Office on Drugs and Crime (UNODC) (2002). *The Bangalore Principles of Judicial Conduct* (United Nations).

Van Brussel S., Timmermans M., Verkoeijen P. and Paas F. (2020). 'Consider the Opposite': Effects of Elaborative Feedback and Correct Answer Feedback on Reducing Confirmation Bias, in *Contemporary Educational Psychology*, n. 60.

Wistrich A. and Guthrie C. (2005). Can Judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding, in *University of Pennsylvania Law Review*, n. 153.

Wistrich, A. and Rachlinski J. (2017). Implicit Bias in Judicial Decision Making: How It Affects Judgment and What Judges Can Do About It, available at *SSRN Electronic Journal*: <https://ssrn.com/abstract=2934295>.

Zenker F. (2021). De-Biasing Legal Factfinders, in F. Zenker (ed.), *Philosophical Foundations of Evidence Law* (Oxford University Press).

Zenker F., Dahlman C. and Sarwar F. (2015). Reliable Debiasing Techniques in Legal Contexts? Weak Signals from a Darker Corner of the Social Science Universe, in *Psychology of Argument*, n. 59.

Zenker F., Dahlman C., Bååth R. and Sarwar F. (2018). Reasons Pro et Contra as a Debiasing Technique in Legal Contexts, in *Psychological Reports*, n. 121.

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

Fictional Mechanical Minds

On the Relationship Between Assumptions of Rationality and Conceptions of Government

PAOLO CAPRIATI

Postdoctoral Research Fellow in Philosophy of Law, University of Palermo (Italy)

✉ paolo.capriati@unipa.it

ORCID <https://orcid.org/0009-0002-6027-441X>

ABSTRACT

If a machine were considered more rational than any other agent, would we entrust it with the government of public affairs? This paper aims to examine the relationship between rationality and political justification. The thesis whose soundness I will verify is: (T1) Depending on which subject's (greater) rationality is assumed, different conceptions of government emerge. The link between assumptions of rationality and political legitimacy implies another thesis: (T) Under certain conditions, to legitimize a subject to decide on matters of collective interest, it is necessary to assume that the deciding subject is rational. Firstly, the terms of the issue must be clarified: (1) the subjects who confer legitimacy on a government; (2) the assumptions of rationality; (3) rationality; (4) the decision-making subjects; (5) machines as decision-making subjects; (6) the conceptions of government. Secondly, under what conditions do the assumptions of rationality become essential to legitimize a government's decision on matters of public interest? Finally, starting from three distinct political subjects – (A) the individuals, (B) the individuals conceived as a collective, and (C) the machine – it will be shown how, by assuming their rationality, three distinct conceptions of government take shape that are: (A1) aggregative democracy; (B1) de-liberative democracy; (C1) machine-government.

Keywords: rationality, political legitimacy, decision-making process, aggregative democracy, deliberative democracy, machine-government

ATHENA

Volume 5.2/2025, pp. 46-70

Articles

ISSN 2724-6299 (Online)

<https://doi.org/10.60923/issn.2724-6299/22550>



1. Introduction – Presentation of the Problem

Let us imagine that it were possible to entrust the government of public affairs to a machine. Why should we do so? There are many possible answers to this question. Let us assume, plausibly, that the following argument is given:

We should entrust the government to a machine because a machine can be more rational than any other agent.

Let us therefore focus on the relationship between rationality and political justification.

The thesis whose soundness I will attempt to verify is the following:

(T₁) Depending on which subject's (greater) rationality is assumed, different conceptions of government emerge.

What I will argue is that there is a link between assumptions of rationality and political legitimacy¹ such that, as the subject whose (greater) rationality is assumed varies, the conception of government also changes.

The link between assumptions of rationality and political legitimacy implies another thesis:

(T) Under certain conditions, to legitimize a subject to decide on matters of collective interest, it is necessary to assume that the subject who decides is rational.

The presentation of these theses is divided into two parts.

¹ "Legitimacy" – understood in a descriptive sense – refers to the mere acceptance of a government by its subjects. "Justification", on the other hand, refers to the reasons used to legitimize a given government. For example, government X is legitimate because the governed hold certain beliefs or a particular faith in it (Weber, 1964); government X is justified because there is a specific reason why the governed regard it as legitimate. The distinction between legitimacy and justification is not as clear-cut as it may seem. More precisely, the purely descriptive concept of legitimacy has been challenged, since it does not consider second-order beliefs about legitimacy – that is, beliefs about what is necessary for a given institution to be considered legitimate. According to Beetham, a "power relationship is not legitimate because people believe in its legitimacy, but because it can be justified in terms of their beliefs" (Beetham, 1991, 11). As will be seen, according to my reconstruction, the reason why, in some cases, the governed consider a government legitimate is that they believe government X to be rational. In other words, under certain conditions, a government is legitimate – and justified – if it is regarded as rational.

In the first part, I will define the key terms of the discussion. First, I will clarify what I mean by “assumptions of rationality” and explain in what sense the law relies on such assumptions for its functioning. Before that, it is necessary to define the concept of rationality itself and specify how it is understood in this context. Secondly, I will clarify what I mean by “conceptions of government”. This notion is closely connected to the model of government and to the political decision-making procedure. I will refer to the political decision-making procedure as the process that leads to decisions on matters of collective interest. It will therefore be necessary to clarify what constitutes such a process, what forms it may take, and how it can be justified. Finally, I will address the possibility that the decision-maker may be a machine, specifying how I use the concept of “machine” and in what sense a machine defined in this way can be considered rational or not.

After defining these essential terms, the second part of the work will investigate the connection between assumptions of rationality and political legitimacy. The aim is to determine the conditions that make it necessary to assume that the political decision-maker is rational. To illustrate the link between assumptions of rationality and political legitimacy, I will present an essential taxonomy in which, as the subject whose rationality is assumed changes, different conceptions of government take shape.

The subjects whose rationality may be assumed, which I will consider, are three:

- (A) individual citizens;
- (B) individuals conceived as a collective (provided that they interact in a certain way);
- (C) machines.

The conceptions of government that arise from these three different subjects are, respectively:

- (A₁) aggregative democracy;
- (B₁) deliberative democracy;
- (C₁) machine-government.

In summary, the two theses I will try to demonstrate are:

(T) In some cases, assumptions of rationality are necessary to legitimate the subject who decides on political matters;

(T₁) In those conceptions of government relying on the criterion of rationality, depending on which subject's rationality is assumed, different conceptions of government emerge.

2. What are the Assumptions of Rationality, and what Purpose do they Serve?

By “assumptions of rationality”, I mean the belief that a certain subject is considered rational. This assumption has been defined as “rationality perfectionism”, and it has been suggested that the law supports this assumption by postulating fictitious cognitive abilities of individuals (Ubertone, 2023). The fictitious nature of these abilities is the subject of recent psychological studies that have shown how the mind works in a way that is far from rational – or at least not always rational (Kahneman, 1994; Stich, 1990).

Rationality perfectionism assumes that the rational subject is the individual. The use of this myth is to some extent necessary for the proper functioning of the law. If we did not assume, for example, that individuals can correctly understand the laws, criminal law would lose its *raison d'être*, since anyone could claim a lack of understanding of the law as a cause for excluding guilt.

The basis of this myth has been challenged, as it relies on the assumption that subjects can act rationally – an assumption that is not universally accepted. The psychological literature of recent years seems to go in the opposite direction. Some works in the social sciences have embraced these studies and have developed governance strategies based on assumptions of mind functioning that are far from rational: one example is nudge. Nudge assumes that individuals are essentially irrational beings (Thaler and

Sunstein, 2008). More precisely, according to the proponents of nudge, an effective way to influence individuals consists in adopting automatic, unconscious, and emotional mechanisms of the human mind – defined by Kahneman as “System 1” thinking – and not relying exclusively on rational and conscious deliberation, which, requiring a certain cognitive effort, is activated less often – defined as “System 2” thinking (Kahneman, 2011).

Ultimately, nudge suggests that the law should adapt to human nature, taking into account cognitive biases and irrational tendencies, to guide choices more effectively, without limiting itself to assuming an ideal rationality that does not exist in reality (Thaler and Sunstein, 2011).

One issue remains unresolved: which subjects’ assumptions of rationality are relevant? The answer seems obvious: they are the same subjects who hold the capacity to legitimize a government. Legitimacy, in this context, is understood as the belief in legitimacy (Weber, 1978, 213) and denotes the adherence or acceptance of the governed toward a given government. In this sense, the subjects who confer legitimacy – as well as those whose assumptions of rationality are relevant – are the governed.

2.1 Models of Government and Conceptions of Government

We have seen how the law’s assumptions of rationality work: it is the law that assumes that members of society are perfectly rational. In the case of political legitimacy, we witness the reverse process: it is the members of society who assume the rationality of those who decide.

It is now appropriate to clarify what is meant by “government”. I assume that the government is the subject, or the set of procedures, responsible for making decisions of collective interest. Such decisions are made through a political decision-making process. “Political”, in this context, can be defined as that which concerns the interests of a given community.

It is therefore necessary to identify the boundaries of the political decision-making process. In ordinary language, we reserve the terms “government” and “political decision” for activities that regulate large groups of people in

potentially all aspects of their lives. However, to understand the basic structure of a political decision-making process, it may be useful to consider a toy example involving a small group of individuals rather than a political community, and a decision of very limited scope rather than a fully political one.

Suppose that in a library, it must be decided whether to close the window or not, and three people are sitting in the library who disagree about what to do. In this case, the decision-making process is what enables them to decide on the window – essentially, whether to leave it open or to close it. The possible decision-making processes, in this case, are at least four:

- (1) A random process, in which the decision is made, for example, by flipping a coin;
- (2) A violent process: after a fight among the three individuals, the strongest decides; here, the decision-making process coincides with physical confrontation;
- (3) An electoral process: the decision is made by voting (possibly preceded by debate), generally following the majority rule. In this case, two votes in favour of “open” or “closed” are enough to decide;
- (4) A process where a qualified subject decides: the decision is delegated to a particular subject, based on a criterion of competence. For instance, the person sitting closest to the window might be chosen to decide, as they know better the effects of the window being open or closed – such as whether insects might enter.

These four are examples of models of political decision-making processes, or models of government. Each could be justified with different reasons. (1) and (2) do not belong to the instruments typically accepted in Western legal-political culture. While (1) could be justified by the idea that chance should decide, (2) could be justified by an appeal to physical strength. (3) and (4), on the other hand, share the idea that rationality might play a role in determining who decides. More precisely, in (3) those present in the library could assume themselves to be sufficiently rational to decide. In (4), they could believe that

among them there is someone capable of making a more rational (or informed) choice than the others – in this case, the person closest to the window. However, (3) could also be indifferent to the principle of rationality and justified solely because, through (3), everyone can take part in the decision, and it is good for everyone to do so.

This brings us to the distinction between “models of government” and “conceptions of government”. By conception of government, I refer to the normative structure that underlies a given model of government. (3), for example, can be understood both as a model of government and as a conception of government, if emphasis is placed on the reasons that justify that particular model. In this sense, for each model of government, we can have as many conceptions of government as there are reasons (or sets of reasons) that provide a normative foundation for that model. For instance, in the case of (3), we could have:

- (A) a conception of government that justifies (3) because it allows everyone to participate in the political process;
- (B) a conception of government that justifies (3) because it assumes that the deciding subjects are rational.

2.2 *On the Concept of Rationality*

An attentive reader will have noticed that I have not yet provided a definition of rationality.² The definition that seems most suitable for our purposes is that of instrumental rationality. According to this view, rationality manifests itself as the use of appropriate means to achieve one’s ends (Kolodny and Brunero, 2013). Therefore, instrumental rationality presents itself as a quality of individuals. “I assume that someone is rational” means that I expect that person to act using the means suitable to achieve a certain end.

² The concept of rationality should not be confused with that of intelligence. Although both can be regarded as qualities of a subject, intelligence is a more demanding concept. To clarify: a thermostat can act in a perfectly rational way – from an instrumental point of view, by regulating the temperature precisely – and yet it seems evident that it cannot be considered an intelligent agent.

Let us try to apply this definition to the case of the library window. In this case, the subject whose rationality could be assumed – that is, the one who could be legitimate to decide – is the subject who makes the decisions appropriate for achieving a given end. But who determines this end? One might object that it is precisely the task of the decision-maker to determine the end – and that this end would coincide with the content of the decision itself. However, framing the problem in this way is misleading. The end, in this case, concerns identifying the best possible state of the window – where “best” is a variable that usually depends on the preferences of those present in the library³ – and not the decision itself. The decision, in fact, is only the instrument through which to achieve the end, namely, determining the best possible state of the window.

2.3 Can we Assume the Rationality of a Machine?

Is it possible to legitimize a machine to decide? If, as we have seen, there is a link between assumptions of rationality and political legitimacy, we must ask whether it makes sense to assume the rationality of a machine. Even before that, it is necessary to clarify what we mean by “machine”.

By machine, I refer to an autonomous non-living agent. I will examine these three aspects separately: (A) agent, (B) non-living, (C) autonomous.

(A) According to Luc Steels (1995), an agent:

- is a system (a set of elements that have relationships with each other and with the environment);⁴
- performs particular functions for another system;
- is able to preserve itself.

It has been observed that not all systems are agents. More precisely, agents must be distinguished from “plant-controller systems”, which:

³ There is a hidden assumption in this statement, namely that the concept of ‘end’ is determined by the desires of those who are present in the room. Probably, this is precisely the appeal of democracy: to communicate to the decision-makers what the group’s end is, thereby enabling them to make the decision that allows that end to be achieved.

⁴ Steels (1995) observes that agents do not necessarily need to be physically situated.

are formed of two essential components, the plant or system whose behaviour is to be controlled, and the controller which measures the state of the plant, and thus aspects of its behaviour, and initiates control actions so as to keep it operating within allowable or acceptable limits of behaviour. Of course, plant-controller systems really have a third component, the environment, but this is usually left in the background in any description of the system (Steels, 1997, 7).

On the other hand, agents – and the environments with which agents interact – shape agent-environment interaction systems (Smithers, 1994). In such systems, there is no plant to control and no controller either. It is the agent that is responsible for initiating and sustaining effective interaction with its environment (Smithers, 1997, 8).

(B) It is necessary to examine the concept of living. Living systems:

are characterized by exergonic metabolism, growth and internal molecular replication, all organized in a closed causal circular process that allows for evolutionary change in the way the circularity is maintained, but not for the loss of the circularity itself. (...) [Cf. Commoner, 1965]. This circular organization constitutes a homeostatic system whose function is to produce and maintain this very same circular organization by determining that the components that specify it be those whose synthesis or maintenance it secures. Furthermore, this circular organization defines a living system as a unit of interactions and is essential for its maintenance as a unit; that which is not in it is external to it or does not exist. The circular organization in which the components that specify it are those whose synthesis or maintenance it secures in a manner such that the product of their functioning is the same functioning organization that produces them, is the living organization (Maturana and Varela, 1992).

For an agent to be considered “non-living”, it must lack even just one of these characteristics.

(C) Greater difficulties in framing arise with the concept of autonomy. Various disciplines – such as biology, law, philosophy, and ethics – have addressed the definition of autonomy. What these definitions have in common is the possibility of “self-law making” (Smithers, 1997). The minimal definition to start from is as follows: an agent is autonomous if it is relatively independent from something else (Steels, 1995). In this case, the independence refers to humans: for a machine to be considered autonomous, it must be independent of its programmer to the extent that its actions are not entirely predetermined by a human agent. This autonomy then translates into the ability to determine its own laws. Phenomenologically, there are two facts indicative of a state of autonomy: the relative impossibility of accounting for behaviours *ex post*, or the relative unpredictability of such behaviours.

Finally, I will not consider intelligence as a useful criterion for defining the machine. In defining the intelligence of a non-living autonomous agent, there are at least three tendencies:

- (1) considering intelligent the machine that exhibits behaviour similar to that of humans (Turing, 1950);
- (2) considering artificial intelligence not comparable to human intelligence, using concepts – such as consciousness (Penrose, 1991);
- (3) considering intelligent the machine whose functioning is not reducible to a mere physical process (Steels, 1995).

Following (1) risks considering intelligent a machine that merely imitates human behaviour. In the case of the chess-playing machine, for example, there are programs carrying out deep searches within the search space, and their impressive performance is therefore no longer regarded as an expression of intelligence (Steels, 1995), but rather as the result of a machine’s ability to draw upon a vast memory. As for (2), in the absence of a clear definition of consciousness, claiming that “intelligence is linked to consciousness” is a fundamentally vague statement. (2), using undefined concepts, excludes *a*

priori that a machine can be intelligent. However, the limit of (3) is that, by drawing such a sharp distinction between cognition and action, it seems to overlook the fact that every cognitive process always originates from a physical process – just as, in the case of the human brain, thought takes shape from neural activity.⁵

In the absence of an appropriate definition, it must be concluded that intelligence is not a criterion that, in this context, can be usefully adopted to describe a machine. Therefore, this criterion must be abandoned.⁶

It has been seen that for the law to fulfil its role, it is necessary to assume the rationality of individuals. But is it possible to also assume the rationality of a machine? This question implies a philosophical query of considerable depth: can machines be rational agents?

This issue is related to the problem of the thinking machine. This problem was first presented by Turing in the famous article “Computing Machinery and Intelligence” in 1950. Since “machine” and “thinking” cannot be defined unambiguously, Turing suggests reformulating the question “can machines think” by replacing it with a less ambiguous question: are there discrete state machines that would perform well in deceiving a human observer, making them believe they are conversing with a human being rather than a machine?

Turing’s reformulation shows that the problem should not be confronted head-on but rather approached indirectly. A lateral approach, like the one proposed in “Computing Machinery and Intelligence”, allows avoiding concepts that cannot be easily operationalised. The merit of this reformulation is also that it makes a response possible. In other words, by providing an empirical standard, this question offers a verifiable criterion based on the imitation of linguistic abilities. In this way, the focus shifts to behaviour (which is observable) rather than a supposed internal state of the machine.

⁵ Unless one intends to adopt a definition of intelligence that is not physically grounded, assuming a form of intelligence that exists independently of matter.

⁶ I have dealt with the topic in greater detail in Capriati (2024, ch. 2).

Therefore, in asking whether a machine can be considered rational, it is appropriate to focus not on internal (and unobservable) criteria but on what can be observed, thus providing an empirical standard and concluding that a machine is rational if it behaves rationally. This “behaviourist” definition of rationality is ultimately the one that matters to those who grant legitimacy to a decision-maker on the condition that they act rationally. In other words, what concerns such subjects is the decision-maker’s behaviour, not the internal processes that determine it.

Russell and Norvig (2016, 40), in their famous “Artificial Intelligence: A Modern Approach”, define a rational agent:

(RA1) For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has.

This definition of rationality corresponds to instrumental rationality. It has the merit of translating the concept of rationality into that of performance (or behaviour): an agent is rational if it maximizes performance in its actions. Rationality is therefore not understood as a particular disposition of the mind (or thought), but as something that can be observed by examining the behaviour of a given agent. (RA1) is not the only definition that Russell and Norvig provide for a rational agent:

(RA2) A rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome (Russell and Norvig, 2016, 4).

(RA3) A rational agent is one that does the right thing (Russell and Norvig, 2016, 39).

With reference to (RA3), it is necessary to clarify what it means to do the right thing. Among the main tasks of moral philosophy is to define what is meant by “the right thing”. According to Russell and Norvig (2016, 39), in the world of AI, a consequentialist approach is commonly adopted, which

evaluates the behaviour of an agent based on the consequences that such behaviour produces. To evaluate the consequences of a certain behaviour, it is necessary to adopt a clear parameter of reference. In other words, assessing whether the consequences produced by agent A are better than those produced by agent B requires specifying the criteria according to which such a judgment is made – better in relation to a defined standard or value system.

This definition makes sense if a well-defined teleological perspective is introduced: if the expected output can be clearly defined, the more rational the agent will be, the more its behaviour allows it to approach that output.

Russell and Norvig (2016, 39) then proceed to distinguish between human rationality and machine rationality:

Humans have desires and preferences of their own, so the notion of rationality as applied to humans has to do with their success in choosing actions that produce sequences of environment states that are desirable from their point of view. Machines, on the other hand, do not have desires and preferences of their own; the performance measure is, initially at least, in the mind of the designer of the machine, or in the mind of the users the machine is designed for.

This distinction does not deserve to be explored further. Regardless of whether it is a human being or a machine, an agent will be considered rational if, based on the objectives it intends to pursue, it acts to maximize the chances of achieving those objectives. At this moment, it is irrelevant to understand how these objectives are determined.

In conclusion and answering the question with which I opened the paragraph, the rationality of a machine can be assumed just as the rationality of any other living can be assumed.

2.4 All the Terms of the Issue

The terms of the issue are now clearer. The thesis “some conceptions of government are based on assumptions of rationality and, depending on the

subject whose (greater) rationality is assumed, different conceptions of government arise” consists of the following elements:

(1) Some subjects (2) assume (3) the rationality of (4) certain others – and (5) machines can, in principle, be candidates to be such agents – and, based on these assumptions, (6) different conceptions of government arise.

(1) The subjects who confer legitimacy on a government, namely, the governed.

(2) The assumptions of rationality consist of the belief that certain subjects are rational.

(3) I refer to an instrumental conception of rationality, which considers rational those subjects who make decisions adequate to achieve certain ends.

(4) The subjects whose rationality is assumed are those who are assumed to be able to make political decisions adequate to achieve a certain end.

(5) Among the subjects whose rationality can be assumed, there are also the machines.

(6) Each conception of government justifies a model of government based on specific reasons.

In the following pages, I will aim to demonstrate how assumptions of rationality determine political legitimacy and to present a brief taxonomy of conceptions of government based on the subjects who are assumed to be (most) rational.

3. Assumptions of Rationality and Political Legitimacy

In this section, I will focus on the relationship between assumptions of rationality and political legitimacy – that is, I will aim to demonstrate (T): how assumptions of rationality politically legitimize a subject to make decisions.

We must, in dealing with (T), understand what conditions make it necessary to consider assumptions of rationality to legitimate a government.

My claim is that it is necessary to assume that governments are rational only insofar as the political system is understood in epistemic terms, that is, insofar as there is the belief that:

- (a) Some decisions are better than others;
- (b) There exists a standard of correctness for decisions that is independent of the government's decision-making process;
- (c) The government is considered the agent capable of discovering what these correct decisions are;
- (d) The government is justified based on (c).

I will refer to this way of understanding political systems as the “epistemic conception” (Capriati, 2024). The reason why, in this case, it is necessary to assume the government's rationality is that, in such a political system conception, the government is politically justified in acting only if it is capable of making the right decision (admitting, beforehand, that there is a right decision and that it is independent of the government's decision-making procedure). In this sense, in order to make the right decision, the government must be rational – that is, it must adopt the appropriate means to achieve predetermined ends.

The connection to justification is straightforward: as stated in (d), the justification of the government depends on its capacity to discover what the correct decisions are. The government that acts rationally – and is therefore capable of discovering the correct decision – is the government considered justified and, consequently, the legitimate government.

Let us assume, for example, that there is a government that makes decisions by reading the coffee grounds at the bottom of a cup. Adopting an epistemic conception of the political system, could such a government be justified? Yes – but only if we were willing to assume (and demonstrate) that coffee provides rationally relevant information with respect to the decisions to be made. Therefore, regardless of the model of government, in a political system based on an epistemic conception, the government's action is considered justified if it is assumed that it acts rationally.

To justify, therefore, the government of a machine, in a political system conceived in epistemic terms, it will be necessary to find that there is a widespread belief that the machine acts rationally.

3.1 Taxonomy of Government Conceptions Based on Rationality Assumptions

In some conceptions of government, the rationality of the subject legitimate to decide is assumed. What I will try to demonstrate is that, depending on the subject whose rationality is assumed, different conceptions of government arise (T_1). This means that the assumption of rationality is an element that determines the conception of government.

I will examine three conceptions of government, showing how, behind each of them, the subject whose rationality is assumed is different. In the case of the classical – or aggregative – democratic conception, the assumed rational subjects are individual citizens, whereas in the case of the deliberative democratic conception, rationality would be the prerogative of the group, understood as a whole. Finally, it will be shown how the possible delegation of decision-making power to a machine is necessarily linked to assuming the rationality of the machine itself.

Before presenting this taxonomy, as we have already seen, it is necessary to recall that by “conceptions of government”, I do not refer to the array of norms and bodies that constitute the institutional apparatus of the state, but to the normative structure adopted to make decisions of public interest.

3.1.1 Aggregative Democracy

When we assume the rationality of individuals, a specific conception of government emerges. The conception of government that assumes individuals are rational, and therefore legitimizes each citizen to participate in decision-making, is democracy in its aggregative form – or the economic theory of democracy.

Jon Elster (1986) states that aggregative theories of democracy see the political process as a means rather than an end in itself. For these theories, the

decisive political act would be a private rather than a public action, namely the individual and secret vote. These theories are thus united by a merely “aggregative” view of the democratic process: the aggregation of individual preferences or interests is the way to achieve a collective social choice. The aggregative version also considers citizens’ preferences as fixed and predetermined. As Elster (1986, 128) says: “It is a market theory of politics, in the sense that the act of voting is a private act similar to that of buying and selling”.

My claim is that, according to this conception of government, those decisions made by aggregating the individual preferences of subjects – whose rationality is assumed – are legitimate. The emblem of this conception of government is the electoral process. Through elections, each subject, equally rational, has equal opportunities to influence the decision.

An objection that may be raised is that, in such a conception of government, it is not necessarily assumed that individuals are rational, but simply that (1) they pursue their own personal interests, and (2) they are legitimate to make decisions because they are the ones who know their interests best. In this sense, individuals’ personal interests are not necessarily rational. However, this objection can be easily overcome. In this context, in fact, the rationality of personal interests is not at issue. Rationality, understood in an instrumental sense, concerns the adequacy of means to ends – which coincide with personal interests – and therefore the ability of individuals to pursue their own interests better than anyone else (or, if necessary, to delegate their expression better than anyone else).

It is clear that in such a system, it is not expected that every decision will always be made by all of the individuals. In practice, there is a division of decision-making according to a principle of competence. Is this mechanism in contradiction with the idea that all individuals are rational? No, since the rationality of each individual is assumed, they are precisely rational enough to recognize they might not be able to decide on everything, realizing that certain decisions would be better if delegated.

In such a conception, therefore, the assumption of rationality concerns individual subjects. The dimension is always collective and is built upon the sum of those individuals.

3.1.2 Deliberative Democracy

From a conceptual point of view, deliberative democracy aims to overcome the aggregative conceptions of democracy (Bächtiger, Dryzek, Mansbridge and Warren, 2018). It can be characterised as that set of conceptions in which public deliberation by free and equal citizens constitutes the heart of the political legitimacy of decision-making and self-governance processes (Bohman, 1998). The distinction between aggregative and deliberative theories is based on the focus of interest: while the former insists on the aggregation of individual preferences – and see voting as the most emblematic representation –, deliberative theories emphasize the transformative possibilities of preferences through dialogical and discursive processes. In the case of deliberative democracy, the subject whose rationality is assumed is the group of individuals. According to this conception of government, rationality is understood as the prerogative of the group rather than of single individuals.

Deliberative democracy has embraced and processed some of the criticisms directed at the idea that the individual is a rational subject. These criticisms are primarily driven by some psychological studies: as seen, these studies have questioned the hypothesis of perfect rationality of the individual (Kahneman, 1994; Stich, 1990). These studies have been accompanied by another literature – also of psychological origin – which argues that certain behaviours, if examined at the individual level, appear dysfunctional, while they may appear functional from a collective point of view (Mercier and Sperber, 2011, 2017; Mercier, 2020). Sperber and Mercier construct an argumentative theory that fits well with some of the main theses of deliberative democracy. More precisely, I refer to the idea that the human tendency to prefer evidence that confirms pre-existing beliefs and to ignore

those that contradict them derives from the ability to construct the best possible argument in support of a particular thesis. This tendency, then, would suggest that human reasoning abilities have evolved to persuade others and take positions in debates, rather than to seek the truth. In this sense, the confirmation bias would be evolutionarily more effective than impartial and disinterested truth-seeking.

The myth of the rational individual is the polemical target not only of psychological studies. Habermas – considered among the founding fathers of deliberative democracy argues –, consistently with the work of Sperber and Mercier, that rationality is a widespread phenomenon and not a quality exclusive to individuals. Moreover, Habermas attempts to overcome the limits of instrumental rationality by proposing a new type of rationality that emerges in a dialogical context and gives rise to the ideal discursive situation (Habermas, 1987). He calls this rationality “communicative” and immediately clarifies that it is not a subjective faculty (Habermas, 2015), but that it takes shape in dialogical and discursive processes.

According to a deliberative conception of democracy, therefore, the subject whose rationality must be assumed is not the individual nor each individual, but the group as a whole and the institutionalised discursive procedure they adopt to reach decisions. Individuals, whose behaviour, if evaluated in itself, can easily be considered irrational, in interacting with each other, would select courses of action informed by rationality.

3.1.3 Machine-Government

Assuming that the group, through interactions among individuals, acts rationally is a fact far from indisputable. Common sense, as well as other psychological studies, seems to go in the opposite direction: in groups, dynamics often arise that negatively affect decisions. Sunstein (2002, 187), for example, observed how deliberation can generate even worse decisions because the “law of polarization” prevails in groups, leading people to cling to predetermined positions.

Questioning the rationality of the group prepares the ground for assuming the rationality of another subject: the machine. I have already anticipated, in paragraph 1.4, the issues related to the possibility of assuming the rationality of an artificial agent. There are no particular reasons preventing the assumption of the rationality of a machine.

No expression explicitly refers to this: even the term “Algocracy” (Danaher, 2016), as it has been defined, does not seem to fit our case.⁷ We could name the conception of government in which a machine is authorised to make decisions of collective interest “machine-government”.

A machine-government is a system in which machines make decisions of collective interest. The hypothesis I am proposing is not that of a machine dictatorship. Rather, “deciding” in this context refers to the mechanism by which inputs coming from human subjects are selected and transformed. In other words, the role of the machine is to collect, weigh, and process individual preferences, with the aim of producing a decision whose justification lies in the idea that the machine assigned to this task is rational and that, more generally, a machine can be a rational entity, or certainly more rational than other agents that could decide.⁸

Some doubts remain regarding the exact determination of the machine-government. I will now present a concrete example.

Delegating public interest decisions to a machine is still a distant reality. However, the Habermas Machine (hereafter HM) (Tessler et al., 2024)

⁷ “I use it to describe a particular kind of governance system, one which is organised and structured on the basis of computer-programmed algorithms. To be more precise, I use it to describe a system in which algorithms are used to collect, collate and organise the data upon which decisions are typically made and to assist in how that data is processed and communicated through the relevant governance system” (Danaher, 2016, 247). In the model I have in mind, the machine does not merely organize the data on which decisions are based, but autonomously makes the decisions itself.

⁸ Rationality cannot be understood as a necessary and sufficient condition for legitimizing a delegation of decision-making power, since it is clear that when assessing the quality of a political decision, one cannot rely solely on the criterion of rationality. A political decision, in fact, to be considered legitimate, requires additional conditions, such as epistemic or ethical ones. Examples may include transparency or the possibility for users’ verification. This analysis, however, explicitly focuses on the criterion of rationality: it is on the basis of this criterion, in fact, that the taxonomy I have just presented is constructed.

represents a concrete realization of a system capable of collecting, processing, and transforming preferences. HM is a project by Google DeepMind based on a system of LLM – large language model (such as Chat GPT). The name “HM” is a tribute to Habermas’s theory of communicative action, according to which when rational subjects deliberate in an ideal discursive situation, they manage to reach an agreement.

HM was designed to improve collective decision-making processes in various fields. For example, it can be used for contract negotiations, conflict resolutions, political discussions, and citizens’ assembly (Tessler et al., 2024, 1).

In these areas, HM acts as a “caucus” mediator. A caucus mediator is defined as one who privately meets each interlocutor before formulating a proposal that can be collectively accepted (Moore, 1987). HM does not merely mediate the discussion but also formulates decisions that are then submitted for approval by the group members. In this sense, the machine acts as a processor and transformer of various instances to shape a decision that best meets the needs of the greatest number of subjects.

4. Conclusions

What is the relationship between rationality assumptions and the legitimacy of political decision-makers? Is it possible to legitimate a machine to make decisions of collective interest?

To answer these questions, I began by illustrating the problem:
Can conceptions of government be distinguished according to the subject whose rationality is assumed?

Firstly, I clarified the elements that compose this question:

- What are the assumptions of rationality, and what purpose do they serve?
- What is a conception of government?
- What is rationality?

- Can the rationality of a machine be assumed?

Through these questions, it emerged that five elements must be taken into consideration:

- (1) The subjects who confer legitimacy on a government;
- (2) The assumptions of rationality;
- (3) Rationality;
- (4) The decision-making subjects;
- (5) Machines as decision-making subjects;
- (6) The conceptions of government.

After having presented all the elements that constitute my research question, I focused on how (and why) assumptions of rationality politically legitimize a subject to decide. Some conceptions of government are based on rationality, that is, on the idea that the decision-maker must inform their action by rationality. These are the conceptions of government that operate within a political system based on the epistemic account. According to these conceptions, therefore, the subject legitimate to decide is the one who decides rationally. Depending on the legitimate subject who decides – that is, the one considered rational – different conceptions of government arise.

More precisely, I examined three conceptions of government (and corresponding rationality assumptions):

- (A₁) aggregative democracy (which assumes the rationality of individuals);
- (B₁) deliberative democracy (which assumes the rationality of the group as a whole);
- (C₁) machine-government (which assumes the rationality of the machine).

In summary, (T) in some conceptions of government, assuming the rationality of the decision-making subject is a necessary requirement to consider such government legitimate, and (T₁) the assumptions of rationality are central in determining the nature of the conception of government.

In this paper, I have addressed the fictional nature – fictional because unsupported by facts – of the political decision-maker. Returning to the

question that opened this contribution: should we entrust the government of public affairs to a machine? If one were to start from an epistemic conception of political systems, the answer could be affirmative, provided that the machine is assumed to be rational. Whether the machine – just like political decision-makers – is actually rational is another matter, one that opens up new avenues for empirical research.

References

- Bächtiger A., Dryzek J. S., Mansbridge J. and Warren M. (2018). Deliberative Democracy: An Introduction, in A. Bächtiger, J. S. Dryzek, J. Mansbridge and M. Warren (eds.), *The Oxford Handbook of Deliberative Democracy* (Oxford University Press), 1-31.
- Beetham D. (1991). *The Legitimation of Power* (Palgrave).
- Bohman J. (1998). Survey article: The coming of age of deliberative democracy, in *Journal of Political Philosophy*, n. 6(4), 400-425.
- Capriati P. (2024). *Macchine che decidono: prospettive di automazione dei processi decisionali in contesti democratici*, PhD thesis, University of Palermo.
- Commoner B. (1965). Biochemical, Biological and Atmospheric Evolution, in *Proceedings of the National Academy of Science*, n. 53, 1183-1194.
- Danaher J. (2016). The threat of algocracy: Reality, resistance and accommodation, in *Philosophy & Technology*, n. 29(3), 245-268.
- Elster J. (1986). The Market and The Forum: Three Varieties of Political Theory, in J. Elster and A. Hylland (eds.), *Foundations of Social Choice Theory* (Cambridge University Press), 103-132.
- Habermas J. (1987). *The Theory of Communicative Action* (Vol. 1) (Beacon Press).
- Habermas J. (2015). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (John Wiley & Sons).

- Kahneman D. (1994). New Challenges to the Rationality Assumption, in *Journal of Institutional and Theoretical Economics*.
- Kahneman D. (2011). *Thinking, Fast and Slow* (Penguin).
- Kolodny N. and Brunero J. (2013). Instrumental rationality, in *Stanford Encyclopedia of Philosophy*.
- Maturana H. R. and Varela F. J. (2012). *Autopoiesis and Cognition: The Realization of the Living* (Vol. 42) (Springer Science & Business Media).
- Mercier H. (2020). *Not Born Yesterday: The Science of Who We Trust and What We Believe* (Princeton University Press).
- Mercier H. and Sperber D. (2011). Why Do Humans Reason? Arguments for an Argumentative Theory, in *Behavioral and Brain Sciences*, n. 34.
- Mercier H. and Sperber D. (2017). *The Enigma of Reason* (Harvard University Press).
- Moore C. W. (1987). The caucus: Private meetings that promote settlement, in *Mediation Q.*, n. 87.
- Penrose R. (1991). The emperor's new mind, in *RSA Journal*, n. 139(5420), pp. 506-514.
- Russell S. J. and Norvig P. (2016). *Artificial Intelligence: A Modern Approach* (Pearson).
- Smithers T. and Moreno A. (1994) (Eds). Notes for the Workshop on the Role of Dynamics and Representation in Adaptive Behaviour and Cognition, 9 and 10 December, Palacio de Miramar, San Sebastian (Spain).
- Smithers T. (1997). Autonomy in Robots and Other Agents, in *Brain and Cognition*, n. 34(1), pp. 88-106.
- Steels L. (1995). When are Robots Intelligent Autonomous Agents?, in *Robotics and Autonomous Systems*, n. 15(1-2), pp. 3-9.
- Stich S. P. (1990). *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation* (MIT Press).
- Sunstein C. R. (2002). The Law of Group Polarization, in *The Journal of Political Philosophy*, n. 10 (2), pp. 175-195.

Tessler M. H., Bakker M. A., Jarrett D., Sheahan H., Chadwick M. J., Koster R., ... and Summerfield C. (2024). AI can help humans find common ground in democratic deliberation, in *Science*, n. 386(6719), eadq2852.

Thaler R., and Sunstein C. R. (2008). *Nudge: Improving Decisions about Health, Wealth and Happiness* (Yale University Press).

Turing A. M. (1950). Computing Machinery and Intelligence, in *Mind*, n. 59, pp. 433-460.

Weber M. (1964). *The Theory of Social and Economic Organization* (Simon and Schuster).

Weber M. (1978). *Economy and society: An outline of interpretive sociology* (University of California Press).

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

When the Nudge Fails: The Limits of Behavioural Individualism and the Case for Meta-nudge

FLAVIO SCUDERI DI MICELI

PhD Student in Human Rights, University of Palermo (Italy)

✉ flavio.scuderidimiceli@unipa.it

 <https://orcid.org/0009-0002-2883-0340>

ABSTRACT

This paper explores the limits of behavioural public policy by critically addressing the debate between i-frame and s-frame approaches. It challenges the dichotomy that frames nudges as tools exclusively aimed at individual decision-making (i-frame), while structural reforms are seen as systemic interventions (s-frame). Drawing on Bobbio's functional taxonomy of legal measures and recent literature on meta-nudges and choice infrastructures, the paper argues that certain behavioural interventions — especially those targeting public officials or institutional processes — can produce systemic effects. Through the example of the “principle of trust” in the Italian Public Contracts Code, the analysis illustrates how regulatory nudges can transform administrative behaviour and reshape decision-making contexts. The aim is to develop a more integrated understanding of behavioural regulation, one that accounts for the interaction between interventions and the social, legal, and institutional environment in which they operate.

Keywords: nudge, normativity and behaviour, influence, behavioural public policy, meta-nudge

I wish to thank the anonymous reviewers for their insightful comments, which greatly improved the quality of the article. I am also grateful to Michele Ubertone and Giuseppe Rocchè for their careful editorial work and valuable suggestions.

ATHENA

Volume 5.2/2025, pp. 71-106

Articles

ISSN 2724-6299 (Online)

<https://doi.org/10.60923/issn.2724-6299/22516>



1. Bounded Rationality in Administrative Behaviour: An Introduction

Traditional economic theories have long described rationality as an optimising choice process, in which individuals select the alternative that maximises their welfare, based on complete information and unlimited computational capacity. One of the first to challenge this paradigm, Herbert A. Simon (1955), introduced the concept of bounded rationality, according to which human decisions are influenced by cognitive limitations, partial information availability, and time pressures that prevent a comprehensive analysis of the available options. As a result, decision-makers adopt *satisficing* strategies, meaning they choose alternatives that are “good enough” rather than optimal. This theory applies particularly strongly to public organisations, which operate within complex regulatory and bureaucratic contexts where rationality cannot be understood as a process of absolute maximisation, but rather as a pragmatic management of the available informational and decisional resources (Simon 1997). Such structures develop routine procedures and standardised decision-making models to reduce uncertainty and ensure internal consistency within administrative processes. However, while these mechanisms may simplify choices, they can also generate rigidity and resistance to change.

In the field of public organisations, a recently established line of research is based on the assumption that the inefficiency or failure of public policies largely stems from the fallibility of human reasoning, and more specifically from behavioural inclinations. From this idea emerged the concept of choice architecture, that is, the context in which we all find ourselves when making decisions. Choice architecture carries the assumption that the context itself influences decisions, and that by altering this context, certain behaviours can be encouraged or discouraged. Since the nudge approach (Thaler and

Sunstein 2008) came to dominate academic debate, efforts have been focused on analysing and developing the components that make up the concept of a nudge, with choice architecture standing as its central pillar.

The public organisation thus becomes the primary decision-making context deserving attention. An administration is a structured system of rules, resources, and actors aimed at pursuing collective goals, within legal constraints and serving the public interest. Unlike private enterprises, which operate under economic efficiency and profit-maximisation criteria, public organisations must balance a plurality of goals – often conflicting – and manage a network of actors with heterogeneous interests. In this scenario, the limitations of human rationality are amplified: public decision-makers not only face fragmented and often contradictory information, but are also bound by regulatory, bureaucratic, and political constraints that affect their choices. As a result, administrations develop standardised procedures to simplify decision-making complexity, reducing uncertainty but also risking inefficiencies and institutional inertia (Crozier 1969). For this reason, if the potential of behavioural sciences applied to law is to be harnessed, public interventions cannot be limited to influencing the decisions of individual citizens, but must also address the internal functioning of institutions, creating mechanisms that foster more coherent and effective decisions. In this respect, one can observe a shift in the concept of rationality within organisations, marking an evolutionary stage in the theory of bounded rationality, where institutions not only operate under rational constraints but also strive to design systems that mitigate the impact of cognitive limitations¹.

¹To be more precise with terminology, one could refer, in Simonian terms, to “institutional rationality” (Simon 1957). Regarding the evolution of the concept of bounded rationality, consideration should be given to the theory of ecological rationality, which holds that rationality is not a matter of following a single universal rule but rather depends on the environment in which the decision is made. The nudge approach, in its foundational assumptions, aligns more closely with the “Biases and Heuristics school” (Kahneman 2012; Stanovich and West 2000). For the perspective of the “Fast and Frugal Heuristics school”, see (Gigerenzer 2000).

This article aims to examine how bounded rationality within administrative organisations influences decision-making processes, and how behavioural sciences can contribute to improving the quality of choices and public contexts. While the nudge approach initially focused primarily on individual decisions, there is now growing interest in strategies that alter rules and decision-making infrastructures, introducing the concept of the meta-nudge. The overarching objective is to demonstrate that the evolution of behavioural policies cannot overlook the cognitive and organisational limitations of the public decision-making process. Only through an integrated approach will it be possible to develop effective strategies to enhance institutional governance and the quality of administrative decisions.

2. The S-frame Critique

The behavioural approach to public policy has frequently adopted an individual-centred model, according to which policy problems stem from cognitive limitations and behavioural biases of individuals and can be addressed through targeted interventions aimed at improving their decisions without altering the underlying rules (Chater and Loewenstein 2023). This paradigm developed from the notion that choice architecture can be modified through nudge strategies that influence human behaviour without imposing constraints, e.g. automatic defaults in pension schemes or nutritional labels to promote healthier eating choices (Thaler, Sunstein and Balz 2014). The effects of these interventions have often been limited and, in some cases, have even diverted attention away from more structural solutions, reinforcing the idea that individuals are solely responsible for their choices, and that political and economic failures can be resolved by acting upon individual preferences without changing systemic conditions.

In response to these considerations, a new perspective has emerged, known as the s-frame, which shifts the focus from improving individual choices to transforming the rules and institutions that shape the decision-making

context. This viewpoint begins with the assumption that many policy problems do not arise from individuals' cognitive errors, but from economic and regulatory structures that condition choices upstream (Connolly, Loewenstein and Chater 2024). From this perspective, truly effective interventions must aim to modify the structural conditions in which decisions are made, for example by regulating the nutritional quality of food products sold in supermarkets rather than merely providing information about their health effects (Chater and Loewenstein 2023,8). Chater and Loewenstein distinguish between types of interventions, defining:

- (1) i-frame (individual frame): an approach to public policy that focuses on individual behaviours and personal choices, aiming to modify them through tools such as nudges, incentives, or information. The central idea of the i-frame is that many social issues stem from cognitive errors or limitations in human rationality, and that these can be corrected without altering the underlying normative or institutional structures of the system.
- (2) s-frame (system frame): an approach that shifts the focus from individual choices to modifying the structures and rules that influence those choices. The s-frame is based on the belief that the primary causes of many social problems are not individual decision-making errors, but rather the institutional, regulatory, and economic contexts in which people make decisions. For this reason, s-frame solutions tend to intervene directly in policies of regulation, taxation, or structural planning to produce systemic change.

One of the main arguments in favour of the s-frame is that the most complex social problems cannot be solved through minor behavioural adjustments, as their causes are often rooted in political and administrative dynamics that require structural change. For instance, while an i-frame intervention to combat climate change might involve providing consumers with information about their carbon emissions or introducing incentives for the use of renewable energy, an s-frame approach would entail the adoption

of carbon taxation policies or stricter regulations on industrial emissions (Chater and Loewenstein 2024). This systemic approach has the advantage of addressing the root causes of the problem, rather than merely mitigating its surface-level manifestations.

Talking about “individual” and “systemic” frames may suggest that systemic effects are simply collective or large-scale effects. That is not what is meant here. “Systemic” does not refer to how many people are affected, but to where the intervention acts in the causal chain of decision-making. An intervention is systemic when it alters the rules, routines, expectations, or infrastructures that organise choices upstream, even if it targets only a limited group of institutional actors rather than the entire population.

The adoption of an s-frame approach may encounter significant political and economic resistance, as it entails deep changes to existing structures and may clash with the entrenched interests of corporations and powerful groups. Sunstein (2022; 2023) and Thaler (2023), on the other hand, have argued that the s-frame risks overlooking the role of individual action and the capacity of i-frame interventions to produce incremental changes that, over time, may contribute to broader transformations. There is also the issue of political feasibility: while nudges are often accepted because they do not impose significant costs on citizens or businesses, s-frame interventions – such as new regulations or taxes – can generate opposition and require greater political consensus to be implemented. The clash between the i-frame and the s-frame is not merely about which policy lever should be pulled first. Rather, it appears to concern the allocation of responsibility for policy failure. I-frame solutions are often promoted as structural substitutes for deeper regulatory reform: by framing social problems as matters of individual choice and self-control, they implicitly shift accountability from institutions and market structures to the individual citizen.

Another criticism that can be raised concerns the risk that the dichotomy between i-frame and s-frame may be overly rigid and simplistic. In many cases, the most effective interventions result from a combination of both

approaches, in which structural changes are complemented by behavioural interventions to support their acceptance and effectiveness. For instance, in the case of social security, an s-frame model based on mandatory pension contributions can be strengthened by i-frame interventions that help individuals better understand their saving needs and manage their financial resources more consciously. Ultimately, the distinction between i-frame and s-frame, while not uncontroversial, should not be entirely dismissed. While early interventions focused on individual corrections, the growing awareness of the limitations of this approach is prompting increased attention to the normative and institutional structures that shape collective choices. This does not mean that i-frames should be abandoned, but rather that they should be integrated into a broader framework, one that seeks not only to improve individual decisions, but also to transform the context in which such decisions are made. The ultimate aim should be to develop policies that not only correct individual behaviours but make virtuous choices the natural outcome of a well-designed system.

The s-frame critique is valuable in drawing greater attention to the effects of interventions on layered contexts, and in shifting perspective on how interventions targeting human behaviour are structured. Indeed, it appears that the general interest of behavioural scientists in individual choice architecture tends to focus predominantly on nudges that work – and work well – such as default rules (Johnson 2022). However, by focusing too heavily on individualised, tailored aspects centred around the person, the true efficacy of the nudge may fail to materialise (Starke and Willemsen 2024). From another angle, the effectiveness of i-frames has been questioned by empirical studies in two key areas: their large-scale impact and their ability to address complex problems. In a study by Della Vigna and Linos (2022), it was found that nudges implemented by governmental “Nudge Units” demonstrated modest average efficacy, with an increase in the take-up (i.e. adoption of the desired behaviour) of just 1.4%. This figure is significantly lower than the average effect of 8.7% observed in published academic studies.

Such a discrepancy suggests that the effectiveness seen in academic settings may not automatically translate when nudges are designed and implemented by government agencies. However, when it comes to complex issues, it has been observed that the actual impact of the i-frame may be modest, even when the intervention appears highly effective on the surface. A recent study on a default rule influencing choices toward renewable energy found that over 85% of Swiss households and 75% of Swiss businesses, despite the higher cost, preferred the green energy option (Liebe, Gewinner and Diekmann 2021). Nevertheless, the final impact of this intervention, despite the high proportion of individuals affected, proved to be minor. The energy system does not respond by instantly producing more green energy for the newly enrolled consumers, and furthermore, such an intervention could not be applied universally as there would not be enough green energy to supply. Thus, while i-frames may work, they often prove to be insufficiently incisive regarding the broader objective that the intervention aims to achieve. This triggers a reinforcement loop in the decision-maker, as the measurement of seemingly positive results may be perceived as effective—even if only modestly so. This belief contributes to a diminished focus on potential systemic interventions and, consequently, to the reallocation of fewer human and financial resources in that direction, thereby reinforcing the maintenance of the status quo (Andreas and Jabakhanji 2023).

3. Towards a Systemic Vision of Behavioural Intervention

We can interpret the distinction between i-frame and s-frame as one that depends on the context to which a given measure is addressed. If we denote with I the intervention (any intentional action – normative, administrative, or behavioural – aimed at modifying the behaviour of one or more individuals) that alters the context, with B the behaviour of the subject (or group of subjects) intentionally targeted by the intervention, and with C the pre-existing context prior to the intervention (i.e., the set of causal factors

determining the subject's behaviour), we can express the relationship as follows: $B = f(I, C)$.

In this function, I represent all relevant aspects of the social control measure that directly or indirectly modify the context C and, consequently, the behaviour. The term I is to be read broadly. It includes not only the formal design of the intervention, but also the strategic intentions, implementation choices and operational constraints of the policymaker or choice architect who introduces it. For our purposes, the choice architect is not treated as an independent variable: what matters is not who designs the intervention, but how the intervention, as implemented, interacts with the context to produce – or fail to produce – behavioural change.

In our analysis, we focus on a single type of actor – the public decision-maker – whom we assume to be constant. The s-frame critique appears to have the insight that the context of a specific choice, C , is composed of two main components, distinct yet interconnected: an individual component C_i , representing the situational and immediate conditions that directly affect individuals' choices and behaviours with respect to that specific decision (including elements such as choice architecture, local incentives, and contingent environmental factors); and a systemic component C_{sy} , which includes the normative, economic, and institutional structures that define the broader framework within which the choice occurs. This latter component provides the rules, constraints, and long-term stability that influence collective behaviour and determine the available options.

It is important to emphasise that the context C is always relative to a specific choice or set of choices. Every causal factor that affects a decision can be classified as belonging to C_i or C_{sy} , depending on its scope and nature. Immediate situational factors that act upon the individual at the time of decision-making fall under C_i (e.g., the order in which options are presented, the scarcity of time in which to decide, or the colour or design of a button that encourages a particular choice). In contrast, structural and collective factors that regulate the broader framework and restrict or expand the decision-

making possibilities are part of C_{sy} (e.g., the absence or presence of electric vehicle charging infrastructure, or a streamlined procedure for authorising renewable energy plant construction). On an aggregate level, a single measure may influence multiple contexts – C_1, C_2, \dots, C_n – each related to a particular choice, thus generating effects across a set of similar decisions. Human beings naturally tend to conform to the status quo (Kahneman, Knetsch, and Thaler 1991). For this reason, we can say that even when C_i is influenced by an effective intervention I , it remains constrained by C_{sy} . Returning to the function: $B = f(I, C_{sy}, C_i)$.

Chater and Loewenstein argue that when an intervention focuses solely on the i-frame, the resulting behavioural change (B) may generate a substitution effect that reduces the pressure for systemic reform. In such cases, C_{sy} remains unchanged in an i-frame-only analysis, since the system itself is not altered in any way. The result is that the impact of such interventions is often weakened, even if a small, visible improvement in behaviour occurs. According to Chater and Loewenstein, the fundamental problem with i-frames is that they can create the illusion of change while leaving the structural architecture intact. If a nudge merely alters individual choices without addressing all components of the decision-making context, the change will be marginal and potentially counterproductive, as it diverts attention away from deeper solutions. For example, if the government introduces nutritional labelling to combat obesity (an i-frame intervention), individuals may be encouraged to make healthier choices. However, this measure does not change the broader food system, which may continue to promote the sale of highly processed and unhealthy foods. If labelling is perceived as a sufficient solution, it may weaken support for more robust s-frame measures, such as a sugar tax or stricter regulations on junk food advertising.

In the analysis proposed by the two authors, the role of behavioural influence and the partial effectiveness of nudges is acknowledged. However, the possibility of influencing the system itself through nudges, as an

alternative to classic command and control interventions, is underestimated. It is certainly useful to adopt a different interpretative lens (Madva, Brownstein and Kelly 2023) when analysing intervention types, but it seems overly hasty to classify the nudge solely as a tool capable of influencing only one component of the context, without generating what they consider to be systemic change. Isn't a nudge that uses a default rule to improve pension plan choices an intervention that alters the pension system?

Adopting a functional model of intervention allows behaviour to be described as the outcome of the interaction between the intervention and the context in which it operates. In this view, the i-frame / s-frame distinction is not an ontological division between different policy instruments, but an indication of which component of the context an intervention primarily modifies: the immediate, situational choice environment (C_i) or the broader institutional and normative structures that organise decision-making (C_{sy}). The familiar assumption in the i-frame / s-frame debate that nudges “belong” to the i-frame, while command-and-control measures “belong” to the s-frame, is therefore misguided. What matters is not the label attached to the tool, but which part of the decision-making environment it durably alters. Downplaying the rigidity of the distinction – as this paper proposes to do – makes it possible to recognise that nudges, under certain conditions, may also have systemic effects, especially when they persistently modify decision-making structures, for example by introducing generalised defaults, standardising procedures, or stabilising new social norms. In this sense, an intervention can be both a nudge and systemic. We will refer to these as “meta-nudges”.

At the same time, structural interventions that disregard behavioural dynamics may prove ineffective or counterproductive. The key lies in developing an integrated vision that considers the interaction between interventions and context, and that can guide the design of policies capable of influencing the deeper dynamics of collective behaviour. To overcome this rigidity of categories, one can envisage an innovative tool that is capable of

acting on the system without resorting to incentives or direct coercion. This type of intervention could therefore address the s-frame critique by using behavioural influence to modify both components of the context identified above. A possible solution for coordinating the two perspectives will be addressed later. It is not the case that only command and control interventions influence C_{sy} and only nudges affect C_i . Both types of interventions concern both types of contexts. A taxonomy of social control interventions offered by N. Bobbio may be useful in illustrating this point.

4. Bobbio's Taxonomy of Policy Interventions

When discussing interventions by public administration, it is important to acknowledge a significant evolution in the role of law and the state. There has been a fundamental shift from a state primarily oriented towards repressing undesirable behaviour to one that, alongside this function, actively promotes socially desirable behaviour (Bobbio, 2007). This transformation is closely linked to the emergence and development of the welfare state, which, in response to new social and economic demands, no longer confines itself to protecting certain interests through the repression of deviant conduct but also aims to encourage innovative and economically beneficial behaviours. Within this context, the traditional technique of negative sanction, used to discourage undesirable conduct, has increasingly been complemented by the positive sanction, such as rewards and incentives, intended to actively promote desired behaviours. This shift deeply affects the functional conception of law, transforming it from a mere instrument of social control into a means of both control and social guidance.

According to Norberto Bobbio, instruments of social regulation can be divided into direct and indirect measures (*Ibidem*, 46). Direct measures attempt to obtain the desired behaviour – or prevent the undesired one – by acting directly upon the behaviour itself, such as using physical force by the police. Included in this category are measures of control and surveillance,

which are primarily negative in nature and aimed at preventing undesirable actions from occurring. Indirect measures, in contrast, aim to influence behaviour by acting on the motivations or conditions underlying the behaviour. Sanctions, as well as facilitations and hindering measures, fall within this category. Facilitation refers to the array of mechanisms through which an organised social group exercises control over its members by promoting behaviours in a desired direction, making their enactment easier or less difficult. Hindering is its direct opposite.

It should be noted that social control measures exist on a continuum, and it is often difficult to identify clear-cut boundaries between categories. To clarify their differences, Bobbio distinguishes three levels according to the causal intensity an intervention seeks to exert on behaviour. These levels can be ordered in terms of how forcefully the intervention acts upon the individual. At the highest level of intensity are (1) constriction or preclusion measures, which aim to bring about the desired behaviour or prevent the undesired one by making it unavoidable or impossible. These include direct measures, such as the use of public force to prevent an action. The objective is to render a particular behaviour necessary (in the case of positive direct measures) or impossible (in the case of negative ones). The second level is occupied by (2) retribution or reparation measures. These are applied after the behaviour has occurred and aim either to attach pleasant consequences to the desired behaviour (rewards), unpleasant consequences to the undesired behaviour (punishments or negative sanctions), or to restore the order disturbed by the behaviour (restorative measures or compensation). Only the latter are considered sanctions in the narrow sense. Finally, there are (3) facilitation and hindering measures, which aim to favour the adoption of desired conduct or discourage undesired conduct. Although they exert a lower level of influence, these measures occupy an intermediate position: like (1) they target the behaviour itself; and like (2), they are indirect in nature, relying on psychological rather than physical pressure.

Traditionally, public administrations have exercised authority primarily through the first two levels of social control, which we may simplify under the term command and control measures, as they necessarily alter costs or impose constraints. In contrast, the distinction between facilitation and hindering appears to mirror the one between nudge and sludge² (Sunstein 2021). Indeed, the nudge may be considered a particular form of facilitation, distinguished by its reliance on cognitive mechanisms. Thaler and Sunstein define a nudge as: “any aspect of choice architecture that alters behavior in a predictable way without forbidding options or significantly changing economic incentives” (Thaler and Sunstein 2021,6). More precisely, a nudge influences the behaviour of “Humans” – individuals who deviate from the assumptions of neoclassical rationality – even though it would be ignored by “Econs”, the rational agents in economic models. Both nudges and facilitation aim to encourage desirable behaviours without imposing obligations or prohibitions. They share a preventive nature, acting before or during a behaviour, rather than as a consequence of it. Their relatively low intensity ensures that freedom of choice is largely preserved: no alternatives are eliminated, but the desired choice is made easier or more attractive.

Facilitation appears to be a broader category than nudge, including interventions that do not necessarily rely on behavioural economics and are often more transparent in their objectives. While some facilitation measures simply aim to remove practical barriers or reduce administrative complexity, others may entail modifications to the economic incentives involved, such as reduced costs, expedited procedures, or material benefits. Nudges, by contrast, are defined precisely by the absence of such changes, they do not significantly alter the payoff structure or impose material constraints. This makes nudging a specific subset of facilitation, characterised not only by its

² The term sludge refers to excessive or unnecessary frictions in decision-making processes – such as paperwork, delays, confusing language, or complex procedures – that hinder people from achieving their goals or accessing services. Introduced by Sunstein (2021), sludge is the negative counterpart to nudges: while nudges aim to facilitate behaviour, sludge impedes it, often unintentionally or through bureaucratic inertia.

cognitive focus but also by the fact that it preserves the existing incentive structure. The distinction is therefore twofold: facilitation may or may not act on incentives, whereas nudges never do. Compared to facilitation, a nudge includes an additional behavioural condition (Congiu and Moscati 2022), it exploits cognitive limits, biases, routines, and habits in both individual and collective decision-making. It does so by embedding these factors into the design of the decision-making environment, what is known as choice architecture. Importantly, a nudge operates independently of prohibiting or adding rationally relevant options, of changing incentives (in terms of time, effort, social or economic sanctions, etc.) and of providing factual information or rational argumentation. What characterises this type of indirect measure is its intrinsic link to the intentional attempt to influence the judgement, the choice or the behaviour of people in a predictable way. In other words, the effect (i.e., the predictable change in behaviour) is a function of the intervention (the nudge), which is itself defined by this intentional attempt (Hansen 2016).

Interventions are always designed and implemented regarding the final effect they are meant to produce. Facilitation typically involves removing practical obstacles or simplifying procedures to make a desired behaviour more accessible, without directly intervening in the decision-making process. Examples include streamlining bureaucratic steps, offering support services, or improving access to information. In Bobbio's terms, these are forms of facilitation: they remove or reduce external barriers that make a desired course of action difficult, costly, or time-consuming. Nudging can be understood as a specific mode of facilitation. Rather than intervening on external obstacles, it works by leveraging predictable psychological mechanisms – for instance by setting defaults, structuring the presentation of options, or reframing outcomes – so as to guide choice without coercion. Both approaches aim to increase the likelihood of certain behaviours, but they act on different fronts. Classical facilitation modifies the external conditions of action, whereas nudging modifies the decision-making context as it is

internally perceived by the individual. For this reason, nudging should not be treated as an alternative to facilitation but as one of its behavioural sub-forms, an indirect measure of social control that operates psychologically rather than materially. The distinction is subtle (and perhaps mainly methodological) but it is analytically valuable. It allows us to treat these indirect techniques within a single family of interventions, while still distinguishing how behavioural change is produced through regulation and clarifying the intentions of those who seek to bring about such change. Bobbio's taxonomy is particularly useful because it shifts the focus from the formal nature of interventions to their functional role in shaping behaviour. Rather than distinguishing measures based on whether they target individuals or institutions, Bobbio classifies them by how they operate, directly or indirectly, and with what intensity. This approach reveals that interventions like facilitation can act not only on individual behaviour but also within institutional contexts, depending on how and where they are applied. In this light, the distinction between i-frame and s-frame does not align with a fixed boundary between "behavioural" and "systemic" tools. What matters is the causal structure of the intervention and its position within the decision-making chain. Bobbio's functional perspective thus helps to overcome the assumption that only traditional rules influence systemic contexts, while behavioural tools are limited to individual effects.

5. What Makes an Intervention Successful or Unsuccessful

Having established the relevant categories for command-and-control and nudge measures, it is essential to distinguish when these interventions are truly efficacious, namely when the public decision-maker's intention to change a behaviour aligns with the actual effect that is produced in reality. All three categories of intervention may be successful or not in influencing either C_i or C_{sy} . Interventions in category (3) may be successful or unsuccessful regardless of whether they impose costs on individuals, and

therefore regardless of whether they qualify as nudges. At least four scenarios must be considered when evaluating the success or failure of an intervention: (A) the intervention generates only the desired behaviour, e.g. a local authority introduces a progressive tariff for waste collection based on the amount produced; as a result, citizens reduce their waste and increase recycling rates; (B) the intervention produces no behavioural change, e.g. the government launches an anti-smoking awareness campaign via TV ads and posters, yet the habits of smokers remain unchanged; (C) the intervention produces the desired behaviour but also causes an undesirable one, e.g. to reduce traffic congestion, a city introduces a congestion charge for access to the city centre, leading to decreased traffic during peak hours but increased pollution in suburban areas (D) the intervention generates only undesirable behaviours – for example, to reduce sugar consumption, the government heavily taxes fizzy drinks; yet, instead of reducing sugar intake, consumers switch to equally unhealthy but untaxed alternatives, such as industrial fruit juices.

Drawing on Tuzet (2016), efficacy can be understood in two distinct ways: in a legal-technical sense, as the ability of an intervention to produce – or at least be able to produce – certain effects; and in a more philosophical sense, as the extent to which the intervention realises the aims for which it was devised. The second definition allows us to further distinguish efficacy from effectiveness, the latter referring to the degree to which an intervention is actually observed or complied with in practice³. These two concepts are independent: one may exist without the other. This means that an intervention can be effective (in terms of implementation) but not efficacious (in achieving its goals), or vice versa. A common belief is that nudges offer public administrations quick, visible results at relatively low political and administrative cost. But this does not mean that they are necessarily

³ In this regard, Tuzet refers specifically to regulatory interventions, that is to legal norms; however, this classification is well suited to broader reflections on the design of any direct or indirect measure.

efficacious in addressing the underlying policy problem. A nudge can display high levels of observed compliance (high effectiveness) and still fail to tackle the structural causes of the issue it targets (low efficacy). It is therefore worth examining how design flaws in an intervention might compromise either its efficacy or its effectiveness. To visually represent the causal relationship between policies and behaviours, we can illustrate an intervention (I) – whether command-and-control or a nudge – with a solid line when the behaviour (a) directly influences behaviour (b). This model corresponds to scenario (A), where both efficacy and effectiveness are beyond doubt. At this stage, it is important to make a preliminary observation: in terms of causal relationships, nudge-based measures and command-and-control measures are equivalent. However, at times, the outcome produced may differ from the one originally intended. Even though the initial intention is to generate a particular effect through the intervention, that specific outcome does not occur, and instead, another effect is produced. This relationship can be represented with a dashed line. This model aptly describes scenario (D), in which only undesirable effects are generated, making the intervention both ineffective and inefficacious. The behaviours initiating the causal chain may or may not qualify as nudges, depending on whether the effect on the next node in the chain is achieved by altering the cost for the individual performing the relevant behaviour.

A further observation on efficacy is necessary. Consider the case in which a government sends personalised letters to taxpayers containing messages such as: “90% of your fellow citizens pay their taxes on time. If you are part of the minority who don’t, you may be fined”. The aim is to encourage individuals to declare their income correctly. The intervention is effective, in that it leads some people to regularise their tax status. However, other citizens, worried about potential consequences, turn to intermediaries (e.g., accountants, consultants), who often continue to suggest evasive strategies, undermining accurate tax calculation. Although the behaviour of some individuals improves, the overall effect of the intervention is inefficacious,

because the broader context continues to offer structural incentives for tax evasion. In this case, the intervention is observed and applied (demonstrating effectiveness), and the intention to modify behaviour aligns with the outcome. Graphically, this would correspond to scenario (C): the intervention is effective to some extent but also generates a directly related undesirable behaviour. If we assess efficacy in terms of the immediate behavioural change achieved, the intervention might appear efficacious. However, if efficacy is judged in relation to the ultimate aim of the intervention – i.e., addressing the systemic problem it was designed to resolve – then it fails to achieve that purpose. Thus, we have two evaluative criteria: efficacy in influencing behaviour and efficacy in achieving the intended target. When the latter is lacking, the intervention cannot be deemed truly efficacious and must be considered a failure. We may now consider scenario (B), in which there is no causal link between fact (a) and fact (b). In such cases, no behavioural change occurs as a result of the intervention. Since the intended outcome is neither enacted nor observed, the intervention is ineffective.

Representation of Causal Relationships Between Policies and Behaviors

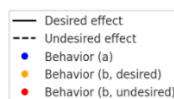
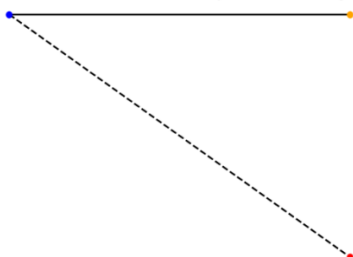
Case (A): Effectiveness and efficacy

Case (B): Ineffectiveness but ultimate efficacy



Case (C): Effective but partial efficacy

Case (D): Undesired effect



Let us suppose, for example, that in an effort to reduce cronyism in the appointment of senior public officials, the government introduces an online portal that publishes in real time the names and CVs of candidates selected for high-level positions. The idea is that increased transparency will discourage non-meritocratic practices. However, the intervention is largely ignored, as entrenched networks of favouritism persist and public officials continue to appoint individuals close to their personal or political circles, regardless of media exposure. Despite this, the heightened attention from the public and the media, caused by the introduction of the policy, increases social and political pressure on the issue, eventually compelling public decision-makers to introduce stricter evaluation criteria for appointments. Thus, even though the original intervention did not function as intended, it still contributed to a positive change aligned with its initial objectives. Cases like this may be rare, but they are not unthinkable. The considerations on efficacy outlined above apply here as well: the intervention is ineffective, yet ultimately efficacious, as it achieves its intended purpose through non-causal means. Nonetheless, the intervention fails, as its implementation was not observed, and the efficacy arose from external factors not inherent in the intervention itself. In scenarios (B) and (C) just discussed, we encounter a common flaw in the design of the intervention. In the first case, the systemic component of the context cancels out the potential individual shift prompted by the nudge. In the second case – such as the role of intermediaries in tax compliance – failure to intervene in the broader institutional conditions results in the emergence of alternative behaviours that neutralise the desired change or even produce unintended effects. In other words, that part of the context (C_{sy}) comprising legal, economic, institutional or social structures remains the same and alters the intended behavioural effect of the intervention.

This taxonomy is relevant not merely for distinguishing between different forms of failure, but because it reveals how both behavioural and normative interventions may fall short when they neglect the systemic component of the context. Nudges without regulatory backing often fail to generate lasting

change, as in the case of personalised tax reminders that do not modify institutional incentives. Conversely, formal regulations that overlook behavioural dynamics (such as complex administrative reforms implemented without adequate support mechanisms) may also prove ineffective. These examples reinforce the need for integrated interventions that address both individual behaviour and the structural conditions under which it occurs, a strategy that the concept of meta-nudge aims to promote.

6. Meta-nudge: Possible Counterexamples to the S-frame Critique?

Across the extensive literature on nudges, numerous authors have proposed taxonomies aimed at classifying interventions based on various criteria: for example, according to the aspect of reality they seek to modify (Münscher, Vetter and Scheuerle 2016), such as information, structure, or decision assistance; the cognitive processes involved (Luo, Li, Soman and Zhao 2023), such as attention, memory, or intrinsic/extrinsic motivation; the medium through which behaviour is influenced, such as digital nudges (Valta and Maier 2025); or even the degree to which individual autonomy is impacted (Baldwin 2014). While these classifications offer useful tools for application, they often fall short of explaining why a particular intervention is efficacious, which causal processes are activated, and which actors are most strongly affected by them. It is possible to move beyond these approaches by analysing nudges through the lens of the causal relationship they establish with the target behaviour. This perspective enables a distinction between two different types of nudges: on one hand, direct nudges, which act immediately on the individual choice; on the other, meta-nudges, which influence the broader context by acting upon intermediate agents (e.g., public officials, financial or commercial intermediaries) who in turn shape the behaviour of others through

mechanisms such as enforcement, normative expectations, or subsequent interventions (Dimant and Shalvi 2022)⁴.

The meta-nudge thus emerges as a form of systemic intervention. It does not aim merely to change individual behaviour, but rather inserts itself into a broader causal chain, acting upon the regulatory and operational conditions that influence the choices of multiple actors simultaneously. It does so by targeting those figures who are in a position to deliberately influence the actions of others. This makes it a particularly powerful tool for public administration, where institutional leverage can be exercised across several levels, generating lasting and scalable effects on collective behavioural dynamics. However, it is important to clarify that the concept of meta-nudge should not be conflated with the broader domain of administrative law or with general measures aimed at regulating the behaviour of public officials. Not all norms targeting intermediaries qualify as meta-nudges. What characterises a meta-nudge is not the normative status of the rule, but its behavioural structure. The intervention must be intentionally designed to influence the decision-making heuristics, perceptions, or framing processes of intermediaries, rather than simply prescribing duties or procedures. In this sense, meta-nudge occupies a specific space within the wider set of systemic interventions, defined by its cognitive and indirect mode of influence. By exploiting the automatic processes of human reasoning, the nudge seeks to create conditions in which the choices of an individual spontaneously align with those intended by the choice architect, without any formal rule requiring such alignment.

This approach evokes the form of governmentalist power described by Foucault as governmentality, i.e. the capacity to create and maintain circumstances in which the actions of a multiplicity of actors naturally

⁴ I will use the term “meta-nudge” to refer to the concept as defined by Dimant and Shalvi (2022). However, the term has also been employed in a different sense, to describe an intervention designed to prompt critical reflection on the effects of another nudge, thereby safeguarding individual autonomy (Gelfand 2023). That said, “counter-nudge” would arguably be a more appropriate term for this latter type of intervention.

converge – without coercion – towards a broader, deemed-optimal order (Brigaglia, 2019). As with the norm-nudges⁵ analysed by Bicchieri and Dimant (2022), the effect of a meta-nudge is not always immediate or linear. Its success depends on the distribution of social expectations, the role played by the actors influenced, and their ability either to amplify or dilute the initial intention of the policymaker. Yet, precisely because of its cascading structure, the meta-nudge represents a theoretical evolution of the traditional nudge, one geared towards the systemic design of behavioural change. If nudges can be linked to a form of governmentalist power, then the meta-nudge appears to embody another modality of influence identified by Foucault, namely a de-subjectified form of power that lies somewhere between intentional influence (power in the strict sense) and unintentional, in other words the anonymous power. In this case, the intervention no longer directly governs behaviour of people, instead operates diffusely and across networks, through repeated actions that follow an originally designed scheme and continue to generate the intended effects, even in the absence of a subject who maintains conscious control over the original intention (Brigaglia 2019,127).

This gives rise to a faceless dynamic of power, in which the initial intervention by the public decision-maker is internalised by other actors – either intermediaries or secondary agents – who then replicate, amplify or institutionalise that influence, often unintentionally. From this perspective, the meta-nudge is no longer merely a technical instrument but becomes the trigger of an impersonal process that fills the space of decision-making with power, without exercising it directly. Although the concept of choice architecture has played a crucial role in shaping the behavioural public policy approach to individual decisions, the growing interest in systemic and multi-level interventions now clearly demonstrates the need to broaden its scope. Interventions limited to specific decision-making moments, however well designed, risk failure when they come up against structural barriers,

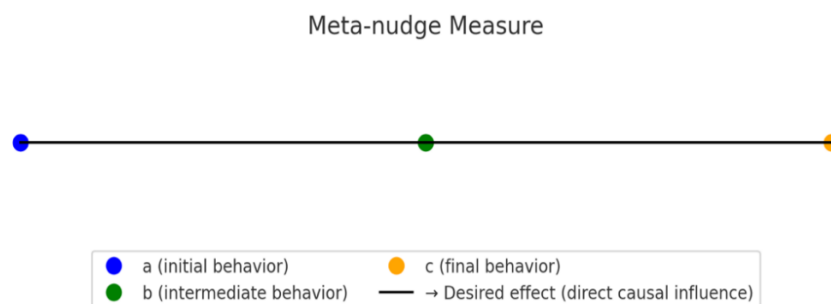
⁵ This type of nudge leverages social norms, i.e., what individuals perceive to be common or desirable behaviour, to steer individual choices.

regulatory misalignments or perverse incentives. Factors that a single choice architecture, by its very nature, is not equipped to address. For this reason, it has been proposed that the notion of choice architecture be complemented – or, in some cases, replaced – by the concept of choice infrastructure (Schmidt 2024). Unlike the former, which focuses on micro-environments of decision-making and targeted interventions on individual behaviour, choice infrastructure refers to the broader systemic conditions, institutional structures, operational processes, and functioning rules that support (or hinder) the efficacy of such interventions over time. It is the invisible yet essential technical framework of the system⁶, the web of causal factors that determines whether a behavioural policy succeeds or fails. Within the context of public administration, this paradigm shift proves particularly valuable for at least three reasons: (1) Decisions do not occur in isolated contexts but within complex and hierarchically organised environments, where citizens, civil servants, policymakers, and intermediaries all interact; (2) institutional dynamics are often slow-moving, distributed and governed by both formal and informal norms, rendering ineffective any approach that fails to account for these systemic constraints; (3) public interventions often aim to influence not only individual but also collective and recurring behaviours, with effects that are spread across multiple categories of actors. In such a context, introducing the concept of choice infrastructure enables the design not only of individual decision points but also of stable and reproducible conditions that facilitate the alignment of desired behaviours with the broader aims of public policy.

This shift in approach is what makes an effective meta-nudge possible: it acts not on a single move, but on the rules of the game themselves, indirectly influencing a wide array of agents and creating a structured pathway for behavioural change. Whereas the “classic” nudge is characterised by a direct

⁶On this point, E. Duflo (2017) offers a particularly fitting metaphor: if choice architecture is like the visible furnishing of a house, then choice infrastructure is the “plumbing” that ensures water flows to the right tap at the right time.

relationship between the intervention and a change in individual behaviour, the meta-nudge introduces a mediated and cascading causal structure, involving multiple actors and decision-making levels. The intention of the public decision-maker – who may be both a legislative or administrative actor – is not aimed directly at the final behaviour, but rather at the design or alteration of the decision-making environment of other subjects, who in turn influence the final behaviour of third persons. The public decision-maker implements a meta-nudge in order to influence an intermediate actor (such as a functionary, manager or organisation). This intermediary then acts as a new choice architect, generating a secondary intervention or reshaping the choice structure for a second subject, who is the ultimate target of the desired behavioural change. Finally, the behaviour of the end-user changes not as a direct response to the original intervention, but because of the action of an intermediary, which was shaped by the meta-nudge.



This causal chain highlights that the meta-nudge is not merely an enhancement of a classical nudge, but rather a form of second-order design, one that acts upon the capacity of other actors to exert behavioural influence. It is here that the systemic dimension becomes evident: the intervention is de-personalised, aiming to generate automatic influence within the intermediary subject. The effect of the intervention propagates through institutional relationships, roles, and implementation mechanisms operating across multiple levels. This means that the original intention tends to give rise to a subsequent intention, one that, even if not identical, is nonetheless oriented

towards producing a convergent final outcome. The concept of meta-nudge is therefore central to the theoretical claim advanced in this paper. By showing that behavioural interventions can operate on the systemic context (not only on individual decision points) meta-nudges directly challenge the idea that nudging is inherently limited to i-frame strategies. They demonstrate that nudging can take on a structural dimension, and thus contribute meaningfully to institutional and policy change. This confirms the central thesis: that nudges, when designed with attention to causal depth and intermediary dynamics, can fully inhabit the s-frame.

6.1 Meta-nudge in Administrative Law: The Principle of Trust in the Italian Public Contracts Code

The introduction of the principle of trust in the Public Contracts Code (Legislative Decree no. 36/2023)⁷ represents an example of a regulatory meta-nudge, that is, a systemic intervention that does not act directly upon the behaviour of citizens or businesses, but rather modifies the conduct of intermediate actors, in this case public officials, who in turn shape the decisions of other involved parties. This kind of intervention, which focuses more on transforming institutional contexts than on influencing individuals, goes beyond both the traditional command-and-control model and the limitations of individual nudges, opening the door to broader reflection on systemic governance.

According to article 2 of Legislative Decree 36/2023, trust becomes one of the founding principles of administrative action in matters of public procurement, alongside those of outcome-orientation, legality, market access, and good faith. This legislative choice marks a decisive shift in paradigm

⁷ The text of the first two paragraphs under examination is reproduced here: “The allocation and exercise of power in the field of public procurement is based on the principle of mutual trust in the lawful, transparent, and proper conduct of the administration, its officials, and economic operators. The principle of trust encourages and enhances the initiative and decision-making autonomy of public officials, with particular regard to assessments and choices concerning the acquisition and execution of services, in accordance with the principle of achieving results.” (my translation).

compared to the previous code, which had been characterised by a defensive and formalistic framework, rooted in a presumption of distrust towards public officials and economic operators. The introduction of the principle of trust is not merely declarative; it aims to bring about a profound cultural and organisational transformation within the administrative apparatus (Carlioni 2024). It is part of a logic aimed at re-functionalising administrative discretion, no longer seen as a grey area to be contained, but rather as an operational space to be preserved through accountability and alignment with public interest objectives. The norm is grounded in a conception of trust that tightly links autonomy and responsibility. Those applying the norm should, on the one hand, have the faculty to assess whether the “conditions of trust” exist in order to depart from mechanical rule application, and on the other, be subject to a form of oversight that prevents arbitrariness and abuse, however without automatically resorting to sanctions that would undermine the very purpose of the rule (Ursi 2024).

From a behavioural perspective, the principle of trust functions as a meta-nudge because it reshapes the expectations, perceptions, and internal constraints of public functionaries and ultimately those of the general public. It re-legitimises the exercise of discretion by reducing the freezing effect caused by what has been referred to as “fear of signing”, i.e. the tendency toward decision paralysis due to fear of future repercussions. The functionary is encouraged to assume an active, diligent, and outcome-oriented role, grounded in a logic of substantive rather than merely procedural accountability. In this way, the administration is no longer simply a constraint on economic activity but becomes an active facilitator for the economic operators with whom it interacts. The legislative intervention does not directly modify the behaviour of citizens but rather reforms the internal operational logic of public administrations, acting upon the intermediate nodes of the public decision-making chain. It is within these figures – functionaries, managers and procedural officers – that the true systemic efficacy of the principle lies.

This cognitive shift in the behaviour of public functionaries is intended to produce, in cascade, a transformation in the behaviour of all actors involved in the administrative procedure. The perception of a more reliable and less punitive system, more oriented towards cooperation than surveillance, encourages greater compliance, reduces litigation, and fosters a higher degree of spontaneous adherence to the rules. The multiplier effect of the principle of trust is realised insofar as it acts upon a class of actors (functionaries) who are capable of structuring the decision-making context of others (economic operators), according to a logic of institutional propagation of influence. This process is all the more efficacious when it is supported by organisational coherence, behavioural training, and reinforcement measures that reward virtuous conduct. It thus represents a form of systemic influence, in which trust operates as a cognitive simplification within decision-making. Its effect on the context is twofold. On the one hand, the principle acts as a transformative factor for the administrative choice infrastructure (C_{sy}), redefining implicit rules and organisational routines, making the system more receptive to change and more outcome oriented. On the other hand, it generates a new form of diffuse normativity, nourished by coherence between regulatory intentions and expected behaviours, giving rise to a networked form of governmentality (Foucault 2004) that does not impose but rather guides, it does not prescribe but instead structures.

This framework fosters distributed institutional learning, where the improvement of processes occurs not only through regulatory means but also through the reinforcement of practices and collective expectations. What is particularly significant about this type of intervention is that it operates without sanctions or formal coercion. Rather than imposing duties or threatening consequences, it seeks to influence the behaviour of public officials by redefining the institutional framework within which their decisions take place. This non-authoritative character is one of its strengths: by avoiding the rigidity of command-and-control mechanisms, it allows for greater adaptability and endogenous appropriation within administrative

practice. As noted in the Italian legal literature, such interventions challenge the traditional conception of administrative power as the exercise of formal legal authority and instead promote a vision of power as intentional influence over decision-making contexts (Zito 2021). In this view, the effectiveness of behavioural governance does not derive from enforceability, but from its ability to embed desirable orientations into the very environment in which decisions are made.

The norm does not alter the costs faced by agents. Its aim is to promote the presumption that the government trusts public officials, thereby reducing the irrational “fear of signing” that affects many public functionaries. At the same time, it seeks to foster among citizens the social norm that officials may legitimately exercise discretion, without modifying the existing legal boundaries of such discretion. Whether interventions of this kind can produce systemic effects remains an empirical question. Conceptually, however, there is no reason to believe they cannot.

This case shows how S-frame interventions can be made through nudging. I-frame and S-frame interventions are not necessarily associated with a specific mode of influence. While the first focuses only on modifying individual behaviours (C_i) and the latter on changing normative structures (C_{sy}) the trust-based meta-nudge operates on both levels (C), activating a circuit of influence between system and behaviour, between rules and practices, between norms and action. The outcome is an administration that perceives itself – and is perceived – not merely as an executor of procedures, but as a relational actor, capable of fostering trust-based and generative decision environments. There is a risk of oversimplifying the discourse by assigning *a priori* superiority to one type of intervention over another, without considering the concrete conditions of design and implementation. The interpretation of the principle of trust as a form of meta-nudge rests on its specific mode of operation. It does not prescribe behaviour, alter legal entitlements, or impose sanctions, but reorients the interpretive and decisional context within which public officials act. What qualifies it as a meta-nudge is

not the presence of regulatory content per se, but the absence of coercion and the presence of a systemic mechanism of indirect behavioural influence. Rather than governing end-users directly, the intervention targets an intermediary class (functionaries) whose practices shape the conditions under which others make decisions. The legislative norm functions as a behavioural lever within institutional practice, aiming to transform defaults, routines, and expectations at multiple levels of administration. This meets the definition of a meta-nudge insofar as it triggers a cascading effect through the reconfiguration of a choice infrastructure (rather than a micro-level choice architecture), generating systemic outcomes through non-coercive, intention-driven influence. While grounded in legal form, its operative logic is behavioural, not prescriptive. For this reason, it can be understood not as classic regulation, but as a paradigmatic case of a norm that governs through trust and systemic propagation rather than command and control.

7. Final Considerations

The claim of this work has been to critically reformulate the conditions under which behavioural interventions in public policy can be considered efficacious, moving beyond static categories and conceptual dichotomies that too often impoverish the analysis. It has sought to question the rigid division between i-frame interventions (targeting individuals) and s-frame interventions (targeting systems), a distinction that tends to assume behavioural tools are inherently incapable of producing structural effects. The critique levelled against the nudge, in its classical formulation, has been valuable in highlighting the risks of depoliticization and systemic inefficacy. However, it implicitly assumes that the form of an intervention determines its capacity to bring about change.

Within this perspective, nudges would be confined to minor behavioural corrections, whereas only hard rules and coercive interventions would be capable of transforming the context. The argumentative path proposed here

aims to show that such a view is inadequate. An intervention succeeds or fails depending on how it is designed, not on the label it carries. The true key lies in the causal structure of the intervention: which behaviours it aims to modify, through which actors, in what context, and with what expected (or unintended) effects. Every regulatory measure – whether legal, incentive-based, or behavioural – can produce systemic effects, provided it is incorporated within a coherent strategy that considers the cognitive, organisational, and institutional constraints faced by the actors involved.

In this sense, the meta-nudge represents both a theoretical and operational proposal capable of overcoming the s-frame critique without abandoning the behavioural approach. By acting on intermediate actors who hold the power to shape the choices of others, the meta-nudge initiates a causal chain that transforms not only behaviours but also the underlying organisational and cultural logics. The example of the principle of trust in the Italian public contracts code clearly illustrates this point: it is a norm that, without imposing, has success in recalibrating administrative discretion, activating the accountability of the functionaries, and fostering cooperation, while reducing bureaucratic distortions. Starting from this example, one can argue that even a nudge, if properly designed, can generate systemic change, and that every intervention should be assessed not by its formal category, but by its causal and relational function.

Thus, nudges can alter social norms and produce systemic effects. It is a mistake to assume that nudges can only ever constitute i-frame interventions. This distinction is crucial: the boundary between i-frame and s-frame does not align with the distinction between “weak” and “strong” tools but rather depends on the strategic objective and the causal network that the intervention is capable of activating. Nudges, when conceived to act upstream of decision-making conditions, can reshape implicit rules, organisational practices, and widely shared social norms. It should be emphasised, that social control measures are not monolithic or mutually exclusive, but form parts of a *continuum*. Coercion, incentives, facilitation and persuasion are all

techniques that follow different but complementary logics. Their efficacy often depends on their capacity to be integrated. In complex and differentiated contexts, a single tool is unlikely to produce lasting outcomes. On the contrary, it is through the simultaneous and well-calibrated adoption of multiple instruments – adapted to the type of actor, the decision-making moment, and the constraint to be addressed – that robust, adaptive, and genuinely transformative public policies can be built. Eventually, the contemporary challenge is not to choose between nudge or rule, or individual or system, but rather to rethink intervention design as intentional action across a multi-level causal chain. Only in this way can we move beyond reductive classifications and build a model of public governance able to generate trust, efficacy and systemic impact.

While this article has focused on the structural, functional and causal dimensions of behavioural interventions – particularly in relation to public governance and the concept of meta-nudge – it is important to acknowledge that such interventions, especially when implemented by public authorities, raise important questions concerning democratic legitimacy and individual autonomy. The debate on libertarian paternalism has long highlighted the potential opacity of behavioural tools and their capacity to influence choices without conscious awareness or deliberative endorsement (Hansen, Jespersen 2013; Baldwin 2014). Although these normative issues fall outside the primary scope of this paper, they remain crucial for evaluating the legitimacy – and not merely the efficacy – of meta-nudging strategies in institutional contexts. Some authors have pointed out that libertarian paternalism, even when well-intentioned, may give rise to subtle forms of asymmetry between those who design the choice architecture and those who are subjected to it. This imbalance, based on expertise, cognitive access, and institutional authority, can result in a condition of “supervised freedom,” where individuals formally retain autonomy but are structurally nudged towards preferred behaviours without fully transparent justification (Galletti, Vida 2018). Future research should therefore complement the functional analysis

of behavioural interventions with a reflection on their normative implications, especially in terms of how they interact with democratic processes and public reasoning, without assuming that such tools are inherently manipulative or illegitimate.

References

- Baldwin R. (2014). From regulation to behaviour change: giving nudge the third degree, *The Modern Law Review*, vol. 77, n. 6.
- Bicchieri C. and Dimant E. (2022). Nudging with care: the risks and benefits of social information, in *Public Choice*, n. 191.
- Bobbio N. (2007). *Dalla struttura alla funzione. Nuovi studi di teoria del diritto* (Laterza).
- Brigaglia M. (2019). *Potere. Una rilettura di Michel Foucault* (Editoriale Scientifica).
- Carloni E. (2024). Verso il paradigma fiduciario? Il principio di fiducia nel nuovo Codice dei contratti e le sue implicazioni, in *Diritto pubblico*, n. 1.
- Chater N. and Loewenstein G. (2023). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray, in *Behavioral and Brain Sciences*, n. 46.
- Congiu L. and Moscati I. (2022). A review of nudges: Definitions, justifications, effectiveness, in *Journal of Economic Surveys*, n. 36.
- Connolly D. J., Loewenstein G. and Chater N. (2024). An s-frame agenda for behavioral public policy research, in *Behavioural Public Policy*, first view.
- Crozier M. (1969). *Il fenomeno burocratico* (Etas Kompas).
- DellaVigna S. and Linos E. (2022). RCTs to Scale: Comprehensive Evidence from Two Nudge Units, in *Econometrica*, vol 90, n. 1.
- Dimant E. and Shalvi S. (2022). Meta-nudging honesty: Past, present, and future of the research frontier, in *Current Opinion in Psychology*, n. 47.
- Duflo E. (2017). The economist as plumber, in *American Economic Review*, vol. 107, n. 5.

- Foucault M. (2004). *Sécurité, territoire, population. Cours au Collège de France (1977-1978)* (Seuil-Gallimard).
- Gelfand S. D. (2023). Nudging, Bullshitting, and the Meta-Nudge, in *Cambridge Quarterly of Healthcare Ethics*, vol. 32, n. 1.
- Gigerenzer G. (2000). *Adaptive Thinking: Rationality in the Real World* (Oxford University Press).
- Galletti G. and Vida S. (2018). *Libertà vigilata. Una critica del paternalismo libertario* (IF Press).
- Hansen P. G. (2016). The Definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove?, in *European Journal of Risk Regulation*, vol. 7, n. 1.
- Hansen P. G. and Jespersen A. M. (2013). Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy, in *European Journal of Risk Regulation*, vol. 4, n. 1.
- Johnson E. J. (2022). *The elements of choice: Why the way we decide matters* (Simon and Schuster).
- Kahneman D. (2012). *Thinking, fast and slow* (Penguin Books).
- Kahneman D., Knetsch J. L. and Thaler R. H. (1991). Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias, in *Journal of Economic Perspectives*, vol. 5, n. 1.
- Liebe U., Gewinner J. and Diekmann A. (2021). Behavioral interventions and climate change: A meta-analysis, in *Ecological Economics*, n. 185.
- Luo Y., Li A., Soman D. and Zhao J. (2023). A meta-analytic cognitive framework of nudge and sludge, in *Royal Society Open Science*, vol. 10, n. 11.
- Madva A., Brownstein M. and Kelly D. (2023). It's always both: Changing individuals requires changing systems and changing systems requires changing individuals, in *Behavioral and Brain Sciences*, n. 46.

- Münscher R., Vetter M. and Scheuerle T. (2016). A review and taxonomy of choice architecture techniques, in *Journal of Behavioral Decision Making*, vol. 29, n. 5.
- Schmidt R. (2024). A model for choice infrastructure: looking beyond choice architecture in Behavioral Public Policy, in *Behavioural Public Policy*, n. 8.
- Simon H. A. (1955). A Behavioral Model of Rational Choice, in *The Quarterly Journal of Economics*, vol. 69, n. 1.
- Simon H. A. (1957). *Models of Man: Social and Rational* (Wiley).
- Simon H. A. (1997). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations* (Simon and Schuster).
- Stanovich K. E. and West R. F. (2000). Individual Differences in Reasoning: Implications for the Rationality Debate?, in *Behavioral and Brain Sciences*, vol. 23, n. 5.
- Starke A. D. and Willemsen M. C. (2024). Psychologically Informed Design of Energy Recommender Systems, in B. Ferwerda, M. Graus, P. Germanakos and M. Tkalčič (eds.), *A Human-Centered Perspective of Intelligent Personalized Environments and Systems* (Springer).
- Sunstein C. R. (2021). *Sludge: What Stops Us from Getting Things Done and What to Do About It* (MIT Press).
- Sunstein C. R. (2022). The rhetoric of reaction redux., in *Behavioural Public Policy*, vol. 7, n. 3.
- Sunstein C. R. (2023). Conspiracy theory, in *Behavioral and Brain Sciences*, n. 46.
- Thaler R. H. (2023). Nudging is being framed, in *Behavioral and Brain Sciences*, n. 46.
- Thaler R. H. and Sunstein C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Penguin).
- Thaler R. H. and Sunstein C. R. (2021). *Nudge: The Final Edition* (Yale University Press).

- Thaler R. H., Sunstein C. R. and Balz, J. P. (2014). Choice architecture, in E. Shafir (ed.), *The Behavioral Foundations of Public Policy* (Princeton University Press).
- Tuzet G. (2016). Effettività, efficacia, efficienza, in *Materiali per una storia della cultura giuridica*, n. 1.
- Ursi R. (2024). La ‘trappola della fiducia’ nel Codice dei contratti pubblici, in *Bilancio comunità persona*, n. 1.
- Valta M. and Maier C. (2025). Digital Nudging: A Systematic Literature Review, Taxonomy, and Future Research Directions, in *SIGMIS Database*, vol. 56, n. 1.
- Zito A. (2021). *La nudge regulation nella teoria giuridica dell’agire amministrativo. Presupposti e limiti del suo utilizzo da parte delle pubbliche amministrazioni* (Editoriale Scientifica).

ATHENA


CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

Law and Surveillance in the Digital Age: The Role of Orientation

MARTA TARONI

Doctor of Philosophy, "Gabriele d'Annunzio" University of Chieti-Pescara (Italy)

✉ marta.taroni2@unibo.it

 <https://orcid.org/0009-0005-6456-5937>

ABSTRACT

Behavioural science-based regulatory techniques are increasingly pervasive across both public and private sectors. The influence strategies employed by digital platforms exemplify a shift toward behavioural governance - one that legitimizes techno-regulation in the name of collective well-being. This paper argues that nudging may serve as a tentative response to online manipulation: policymakers can deploy counter-nudges to resist the behavioural tactics of digital giants and promote more autonomous decision-making. At the same time, this paper explores how the philosophy of orientation can offer individuals tools to cultivate awareness, recognise influence, and preserve autonomy in algorithmically mediated environments. Taken together, these two approaches - institutional nudging and individual orientation - may work in tandem to support agency and in-formed decision-making. The paper concludes by suggesting that everyday practices of attention and routine can serve as subtle forms of resistance in an age increasingly defined by digital disorientation.

Keywords: behavioural sciences, nudging, philosophy of orientation, disorientation, practises of self, pre-commitment

ATHENA

Volume 5.2/2025, pp. 107-141

Articles

ISSN 2724-6299 (Online)

<https://doi.org/10.60923/issn.2724-6299/22490>



*“The real problem of humanity is the following: we have
palaeolithic emotions, medieval institutions, and god-like
technologies.”*

E.O. Wilson

*“[...] the art of living is first of all a clever art of
orientation.”*

W. Stegmaier

1. Introduction

Today, regulatory techniques grounded in behavioural sciences are increasingly pervasive across both public and private sectors. The behavioural influence strategies deployed by digital giants are emblematic of this trend. They embody a behavioural-governance turn within liberal democracies, providing moral cover for techno-regulations steering individual choices while promising collective well-being.

In this paper, I will argue that nudging can be used as a tentative solution to the problem of online manipulation, both individually and collectively: policy makers and institutions could use nudging techniques, as a counternudge, that is to contrast the behavioural influence strategies put in place by digital giants and to channel behaviour toward greater autonomy; individuals, through the *philosophy of orientation* (Stegmaier, 2023; Stegmaier, 2019), should cultivate self-awareness and decision-making capacities that help people recognize manipulation and preserve agency. Taken together, these approaches suggest a way to reorient regulation in an age of disinformation and crisis for traditional legal frameworks.

2. The Evolution of the Decision-making Agent Model and New Forms of Normativity

Firstly, I would turn to the evolution of the decision-making agent model, beginning with the conception offered by classical liberal theory. Here, the subject is imagined as one who has emancipated himself from the “state of minority”, standing free among others and equal. Within this *Weltanschauung*, citizens are understood – borrowing Rawlsian terms – as both rational and reasonable, capable of articulating and pursuing their own comprehensive life-plans. The role of state institutions is thus circumscribed: they are charged with demarcating the boundary between legitimate public action and the private domain beyond which intervention is impermissible. The liberal legislator bears the determinate task of safeguarding this sphere of autonomy, the very condition of possibility through which subjects realise their chosen ends. Institutional action proceeds under the presumption of individuals as fully capable agents – able to discern what is best for themselves, to cooperate with others, and to negotiate with the State to secure their place within the social order. The proclaimed neutrality of law and politics vis-à-vis the plurality of life-plans chosen by rational and autonomous individuals is what guarantees the sovereignty of the liberal subject as *homo oeconomicus*. This assumption has been increasingly criticised and challenged by the post-liberal legal theories (Minda, 1994), which have emerged since the late twentieth century and seek to interrogate the very ontological assumptions on which classical liberalism is founded. Theorists of caring, for instance, argue that the capacity for self-determination – the ability to design a free life plan – is in fact merely a formal faculty of the decision-maker. In a coeval and parallel intellectual trend, investigations into cognitive bias and bounded rationality have proliferated, giving rise to the heuristics analyzed by the behavioural sciences (Simon, 1955, 99-188; Simon, 1982; Kahneman, 2003, 1449-1475; Kahneman, 2011; Kahneman and Tversky, 2011, 453-458; Kahneman and Tversky, 2000). These studies

underscore the necessity of rethinking the relation between subject and norm no longer as an idealized abstraction, but in a more phenomenologically grounded and realistic manner. What is at issue is not only the production of obedience to rules, but rather the spontaneous internalization of norms and the normalization of conduct. The critical thrust of post-liberal jurisprudence lies precisely here: liberal individualism is predicated upon idealized conditions of agency, abstracted from the finitude and vulnerability of concrete human subjects. In the liberal imaginary, the legislator presupposes a normative addressee who is free, reasonable, rational, prudent, skillful, even quasi-omniscient – perfectly capable of discerning and actualizing the dictates of the law.

More recently, within the post-liberal context, behavioural economic theorists (Kahneman, Thaler) have challenged the model of the rational “Econ”, replacing it with the more realistic “Human”. Humans are not free and equal decision-makers but individuals shaped by necessity, gender, birthplace, economic constraints, and education. They are formally free to choose among options, yet their rationality is context-bound, partial, and often flawed. Once rationality is recognized as biased or limited, individuals can no longer be seen as fully autonomous in the liberal sense; rather, they are agents of bounded rationality whose errors can have significant personal and societal costs.

This recognition has prompted a rethinking of regulation. If individuals are prone to systematic mistakes – whether in health, finance, or everyday choices – public institutions may need to account for these vulnerabilities. Here lies the delicate boundary between regulating conduct and merely influencing it, between predicting behaviour and normatively evaluating it. Sunstein and Thaler’s theory of nudges offers a thoughtful response to this question (Sunstein and Thaler 2003, 1159–1202; Sunstein and Thaler 2008; Sunstein 2013; Sunstein 2014). Through “choice architectures”, regulators can steer individuals toward better outcomes without eliminating freedom of choice. Nudges rely on small, targeted interventions, empirically grounded in

behavioural economics, to produce substantial effects in both public policy and private domains. This approach embodies “libertarian paternalism”: a light, non-intrusive paternalism that seeks to reconcile individual freedom with protective guidance (Sunstein and Thaler, 2008).

Presupposing a decision-maker affected by biases and reasoning flaws – and thus moving away, at least partially, from the traditional liberal model in which the addressee of the norm is a rational agent always capable of choosing what is best – opens the way to significant benefits for both individuals and society. Sunstein and Thaler argue that humans often need to be “nudged” when facing complex decisions, especially in situations where feedback is inadequate or information is scarce, making it difficult to correctly process environmental stimuli. In such cases, nudges provide a subtle form of assistance.

It is crucial, however, that nudges remain non-coercive. A genuine nudge must leave individuals free to ignore it and make a different choice if they wish. Its purpose is not to impose, but to act as a light prod, a gentle push. The ambition of Sunstein and Thaler is to equip both public and private actors with effective tools that can, on one side, address collective challenges that burden society as a whole, and, on the other, improve the lives of individuals by helping them navigate the maze of cognitive traps. In this view, nudging techniques are not an alternative to policy but a way to reinvigorate it, offering new strategies after decades of repeated failures in markets and social policies.¹ For Sunstein and Thaler, one of the main obstacles to effective governance lies in bureaucratic inertia and the lack of scientific literacy within public administrations. In particular, the insufficient understanding of how the human mind actually works undermines the efficiency and effectiveness of state action itself (Sunstein, 2013a).

¹ In recent decades, several governments have institutionalized nudging through dedicated *Nudge Units*. The most prominent examples are the U.S. Office of Information and Regulatory Affairs (OIRA), directed by Sunstein from 2009 to 2012 under President Obama, and the U.K. Behavioural Insights Team, established by Prime Minister Cameron. Both have applied behavioural economics to public policy — from health campaigns to energy saving and consumer protection — seeking more effective and cost-efficient regulation.

In their work, Sunstein and Thaler draw on cognitive psychology and behavioural economics to show that individuals often make suboptimal decisions due to limits of attention, information, cognitive capacity, and self-control. This insight underpins the theory of nudging: when faced with complex choices, decision-makers are vulnerable to systematic biases, errors of judgment, and weaknesses of will. Sunstein and Thaler argue that rational choice theory rests on a false assumption: that people always act optimally. In reality, sound decisions are possible only when individuals possess sufficient knowledge and experience in context. Moreover, they dismantle the idea that paternalism necessarily implies coercion. With the notion of “libertarian paternalism”, they propose a soft form of guidance that seeks to improve people’s choices while preserving freedom. Unlike coercive paternalism, which restricts autonomy, this approach respects individual liberty by leaving room to opt out of interventions. Although the term may appear paradoxical, Sunstein uses it to challenge the negative connotations of paternalism and to show that paternal guidance and individual freedom can coexist. Within behavioural economics, nudging thus occupies a space between coercion and complete state inaction. It illustrates a middle ground in which institutions can encourage choices that promote individual well-being without undermining the autonomy central to the liberal tradition (Sunstein and Thaler, 2008, 5).

According to Sunstein, nudging techniques operate within this intermediate space, where influences on individual choices are in many respects unavoidable. Endorsing a libertarian framework does not mean rejecting all forms of paternalism: in certain cases, restrictions on choice may be necessary, particularly when constitutional rights are at stake. Yet Sunstein insists that such interventions must always be evidence-based, grounded in empirical studies that observe how people actually behave in daily life. This knowledge allows policymakers and private actors alike to anticipate responses and design more effective solutions. For this reason, supporters of nudges argue that a strictly anti-paternalistic stance is neither coherent nor

workable. Since individuals are prone to systematic cognitive errors, merely providing information is insufficient; what matters is how information is presented, and presentation itself can never be neutral.

As already emphasized, traditional economic theory, by contrast, rests on the image of *homo oeconomicus*: a rational, calculating agent endowed with stable preferences, strong willpower, and flawless computational ability. Such a being – the “Econ” in Sunstein’s terminology – is an ideal type, immune to mistakes in judgment or self-control. Econs belong to an abstract “Econworld”, where choices are processed exclusively through slow, deliberate “System 2”² thinking and where cognitive biases simply do not arise. Humans, however, do not inhabit Econworld. Unlike Econs, they are prone to predictable and recurrent errors, and nudges can therefore exert a powerful influence on their behaviour. In a society composed only of Econs, nudges and libertarian paternalism would be redundant. But behavioural economics shows convincingly that our world is not populated by idealized rational actors; it is populated by fallible human beings, prone to mistakes in prediction and planning, and often struggling with problems of self-control.

In fact, Thaler and Sunstein explicitly challenge the neoclassical model of rational behaviour. Behavioural economics, supported by empirical research such as prospect theory, has shown that human choices reflect bounded rationality, limited self-interest, and limited willpower. Building on these findings, they develop tools for law and public policy. If cognitive biases are systematic and predictable, a well-designed nudge can be used to counteract them for the benefit of both individuals and society. To illustrate this, Sunstein and Thaler look at the free market as a setting where nudges may work effectively. In some cases, competition and consumer protection laws sanction companies that exploit cognitive biases or disseminate misleading information, thereby discouraging overtly unfair practices and fostering a minimum standard of transparency in market interactions. However, such

² For Kahneman, we have two modes of thought: “System 1” is fast, instinctive and emotional; “System 2” is slower, more deliberative, and more logical.

legal safeguards are often reactive rather than preventive, addressing individual violations rather than systemic patterns of manipulation. Moreover, market incentives alone are not always sufficient to ensure ethical conduct. Firms that refrain from exploiting human cognitive limitations may find themselves at a competitive disadvantage, gradually displaced by those that deliberately deploy behavioural influence techniques to capture attention, manipulate choices, and maximize profit. This dynamic produces a paradox: the same competitive mechanisms, that should reward efficiency and innovation, instead end up reinforcing manipulative strategies, eroding both consumer autonomy and market integrity. In this scenario, it is crucial for all economic actors to understand the lessons of behavioural economics – not only to guard against exploitation but also to turn insights about human behaviour toward positive ends (Sunstein, 2008, 7; Sunstein, 2013b, 1832). Within this framework, System 1 thinking is associated with Humans, while Econs rely on the more deliberate System 2. Sunstein likens Humans to Homer Simpson – fallible, impulsive, and error-prone – whereas Econs resemble the hyper-rational Mr. Spock. If System 1 generates biases and System 2 has the potential to correct them, nudges can support the reflective system by subtly steering the automatic one, without coercion. Nudge theory is built on this dual-system model: the automatic system introduces distortions, while the reflective system seeks to correct them. At the heart of this approach lies *choice architecture*. A nudge functions through the design of an environment that guides decisions in beneficial directions. Policymakers, doctors, marketers, and even parents act as choice architects whenever they shape the context in which others decide. Thaler offers a simple example: students leaving a classroom faced doors with large wooden handles. The handles implicitly suggested pulling, while the situation called for pushing. Here, the Automatic System dominated, swayed by the physical cue, even against rational reflection. For Thaler, the lesson is clear: environments must be designed to align signals with intended actions. Good

choice architecture avoids conflicting cues and instead provides clear, consistent prompts that make desirable behaviour easier and more intuitive.

Because our world is populated by Humans rather than flawless Econs, social environments must be designed with real cognitive limitations in mind. Daily life confronts individuals with countless decisions and stimuli, and even small details in design can make a decisive difference. Nudges, therefore, exploit the fact that in the human world – unlike Econworld – minor changes in context can profoundly shape behaviour.

Decades of behavioural research have demonstrated that we inhabit a world already structured by predefined frameworks of orientation – frameworks that subtly yet powerfully shape our decisions. At times, these frameworks guide us toward beneficial outcomes, but they can equally be deployed for harmful purposes, with significant repercussions at both the individual and collective levels.

In the present age of surveillance capitalism,³ the task of maintaining orientation in a world saturated with hidden architectures of influence has become a compelling and extraordinarily challenging one. Every day, individuals are confronted with increasingly sophisticated techniques designed not only to capture attention but also to disorient and to reorient choices in accordance with the objectives of powerful digital platforms and those who command their infrastructures. In this sense, the effort to “find one’s bearings” is continually destabilized and reconfigured by invisible architectures of influence.

It is in this context that legal regulation faces one of its most profound crises. Traditional frameworks of law and regulation, built on the assumption of autonomous and rational subjects, are increasingly inadequate when

³ I will elaborate on the notion of “surveillance capitalism” in the following section; however, it is already important to underline that it refers to an economic order based on the extraction, prediction, and commercialization of behavioural data. Since Zuboff coined the term “surveillance capitalism” in 2014, it has been taken up by legal scholars such as Cohen, media theorists like Couldry and Mejias, and philosophers including Han, becoming a key category in debates on the digital economy (Zuboff, 2014; Zuboff, 2019a; Zuboff, 2019b; Cohen, 2019; Couldry and Mejias, 2019; Han, 2017).

confronted with the pervasive and opaque use of behavioural influence techniques by private actors. In the following sections, I will examine in greater detail the structural shortcomings of classical legal theory – and of policy-making – in the digital age and the urgent need for new normative approaches. For the moment, however, it is sufficient to emphasize that in an era of continuous digital surveillance, we run the risk of losing orientation without even being aware of it – a disorientation that is all the more dangerous because it operates silently, beneath the threshold of conscious awareness.

3. The Loss of Individual Orientation in the Behavioural Influence Society

Methods aimed at influencing behaviour are widespread in the private sector. They are becoming increasingly pervasive in everyday life through digital spaces and algorithmic technologies. A concrete application of behavioural science research can be seen in techniques of behavioural influence, which constitute the very foundation of what Zuboff defines as “surveillance capitalism” (Zuboff, 2019a, Zuboff, 2019b).⁴ A decisive factor in the functioning of this industry is the systematic use of methods derived from cognitive and behavioural sciences to induce users to: 1) devote as much time and attention as possible to the content offered by platforms; 2) disclose personal data; 3) purchase goods or services promoted by advertisers; 4) sustain online and offline behaviours – whether by commission or omission – that conform to the interests of advertisers or of the platform itself. It can be said that the digital market is largely structured around the production and sale of this very power of influence.

⁴ It is important to note that Zuboff’s *Surveillance Capitalism* has generated a copious body of critical responses. For example, both Morozov and Doctorow dispute her central claims: Morozov criticises the work for its overblown and weakly substantiated narrative, while Doctorow reframes the problem as one of monopolistic market power rather than the irresistible force of behavioural surveillance (Morozov, 2019; Doctorow, 2021).

But this exercise of power over individuals can generate a wide range of negative consequences, both at the personal and at the collective level. In some cases, these negative effects are not collateral but intrinsic to the business model itself: profit extraction would be impossible without them. Two main categories of damage can be identified: individual harm, in the form of detrimental effects on users' physical and mental health; and social harm, in the form of the erosion and compromise of democratic institutions. Yet both categories of harm currently appear, at least in broad terms, to be largely ignored by the law. To a reasonable degree of approximation, it can be said that the digital market fundamentally revolves around the creation and commercialization of a specific commodity: the "power of influence". When economic actors in this sector harness behavioural data to shape and predict user conduct, their profits grow exponentially. Without the capacity to generate behavioural predictions, a company such as Google would have no product to sell. The actual commodities Google offers to its paying customers are "predictive products": probabilistic models of individual behaviour, sold to advertisers eager to optimize the effectiveness of their campaigns.

And this distinctive market is sustained by the traces users leave while navigating digital environments. These traces – search queries, reading patterns, purchase histories, geolocation data, and more – are meticulously harvested and stored as data. Such data function as repositories of individual activity, which are then analyzed, processed, and converted into predictive insights. These insights are sold to advertisers, who use them to tailor marketing strategies, deliver highly targeted content, and maximize returns on investment. Behavioural data have thus become the cornerstone of a new business model: an economy structured on surveillance capitalism. Vast volumes of personal information are systematically processed by advanced AI systems with a single overriding aim: the maximization of advertising profits. Every online interaction, often unbeknownst to the user, engages them in a hidden economic transaction. The ostensibly "free" services provided by digital platforms are financed by advertisers, who pay for privileged access

to user attention and behavioural forecasts. To maximize the value of this investment, platforms strive to align user behaviour with the commercial objectives of advertisers.

This alignment is achieved through algorithmic techniques designed both to extract data and to predict future conduct. The collected data become the raw material for predictive algorithms, which are engineered to anticipate user behaviour and subtly steer it toward outcomes favourable to economic operators, and the entire system works thanks to cutting-edge algorithmic technologies capable of large-scale monitoring, fine-grained control, and the systematic exploitation of personal data. In this sense, the ambitions of the digital giants are expansive: they aspire to build exhaustive profiles of each individual, and user cooperation – willing or unwitting – is indispensable in generating the torrents of data that enable algorithms to know us, to predict our actions, and, when deemed profitable, to influence them. Algorithms, drawing on colossal datasets organized hierarchically, perform increasingly precise calculations to anticipate interests, preferences, and purchasing intentions.

This distinctive business model, briefly outlined here, has been aptly labelled “surveillance capitalism”, an economic order that capitalizes on human experiences as raw material for practices centred around data extraction, prediction, and commercialization. The success of this new capitalism hinges on several key factors: 1) the rapid integration of the internet into our daily lives; 2) the discovery of the internet as a reservoir for collecting valuable user actions; 3) the transformation of these actions into behavioural data; 4) the utilization of this data for commercial, often manipulative, purposes. In this economic system, the production of goods and services is subordinate to a global architecture of behaviour change. Surveillance capitalists transform human experiences into behavioural data, with some used to enhance products and services, while the remainder becomes the private behavioural surplus. This surplus is instrumental for the creation of predictive products through the application of artificial

intelligence technologies, capable of forecasting our actions, preferences, and purchases. These predictive products are then traded in a burgeoning market for behavioural predictions. In this evolving economic structure, the market for future behaviour yields substantial profits for capitalists. It is a market that extends beyond the virtual realm, with online behavioural data mining translating into tangible offline consequences. The instrumentalizing power that emerges from this economic system is a pervasive force that understands and directs human behaviour toward specific objectives, insinuating itself into every facet of daily life.

To sustain the relentless growth of surveillance capitalism, its operators are compelled to continually augment their instrumentalizing power and means of behaviour modification. Just as industrial capitalists had to perpetually enhance the means of production, surveillance capitalists must evolve their methods to extract and modify human experience. The digital connection is not the ultimate objective of this new capitalism; rather, it serves as the conduit for a cyclical process that commences with the extraction of human experience and culminates in the modification of behaviour.

The power I have termed “influence” lies at the core of the Internet economy. It represents a unique form of non-normative, non-violent, and non-coercive power. However, it diverges from mere “persuasion”. According to standard definitions, “persuasion” involves convincing or inducing a person to acknowledge the reality of a fact or the validity of a particular state of affairs, typically through the presentation of good reasons or effective communication. In particular, according to the *Cambridge English Dictionary*, persuasion means “to make someone do or believe something by giving them a good reason to do it or by talking to that person and making them believe it”. Persuasion usually operates through language, whether verbal or non-verbal, and relies on the provision of epistemic or prudential reasons for believing in the truth of specific propositions or engaging in certain behaviours. It presupposes the subject’s awareness of being

persuaded. In contrast, the conditioning power underpinning the internet market is generally effective because it remains concealed from the conditioned subject. Influence operates through causal processes that bypass explicit communication of reasons for believing or acting and are, in fact, incompatible with it. As Zuboff discusses (Zuboff, 2019b, 199-233), its effects extend unknowingly to our offline lives.⁵

In the early stages of surveillance capitalism, data extraction was primarily concerned with individuals' virtual conduct. In recent years, however, the focus of major technology corporations has shifted toward what is often described as the "business of reality". This encompasses a wide range of developments – including "environmental computing", "ubiquitous computing", and the "Internet of Things" (IoT) – all of which enable increasingly accurate forecasts and, consequently, more effective strategies of personalized advertising. Smart devices such as Amazon's Echo and Google Home act as sensors embedded in the physical world, processing hours of informal dialogue in order to generate advanced predictions that anticipate user needs.

Because data serves as the lifeblood of this emerging market, technology companies have devised increasingly sophisticated tools to obtain a deep and detailed understanding of users. Through the relentless capture of attention and pervasive forms of surveillance, they have succeeded in constructing

⁵ A still compelling, though somewhat dated, example of this business model's intrusion into everyday reality is the mobile game *Pokémon Go*, developed by Niantic Labs (a subsidiary of Google). *Pokémon Go* is a free-to-play video game, available on iOS and Android, which leverages geolocalized augmented reality with GPS. To progress in the game, players must physically move around, using their smartphones to capture Pokémon. These creatures are often strategically placed in locations established by advertisers, such as McDonald's restaurants and Starbucks coffee shops. The game's true objective is not solely to provide entertainment but to incentivize players to consume in those establishments. The game itself is free for players, and Niantic Labs derives its revenue from advertisers who pay to have PokéStops and Pokémon strategically placed in their desired locations. The effectiveness of this *business model hinges on its covert nature*. Moreover, smart devices that act as sensors within the material world – such as Amazon's Echo or Google Home – process hours of informal dialogue to generate advanced predictions that anticipate our needs (Middleton, 2016; Revesz, 2016).

highly precise predictive models of human behaviour. Importantly, surveillance and behavioural manipulation are no longer confined to computers, tablets, and smartphones; they extend across a growing ecosystem of “smart” products. Wearables and voice assistants, equipped with advanced sensors, not only assist users but also analyze their actions, thereby producing predictive insights.

Profit generation in surveillance capitalism depends on the active steering of behaviour. The same methods that influence consumer decisions regarding products and services can be redeployed to shape other domains of choice. This adaptability has already been demonstrated in politics, particularly in twenty-first-century electoral campaigns. The involvement of Google founder Eric Schmidt in President Obama’s 2008 campaign exemplifies the extent to which technology companies possess the capacity to shape decision-making processes across a variety of contexts.

4. The Power of Influence

This *power of influence* denotes the deliberate shaping of real-world behaviour through techniques grounded in behaviourist psychology. To sum up, these methods encompass: 1) conditioning (shaping behaviour through reinforcement); 2) exploitation of immediate gratification (capitalizing on humans’ inclination toward instant rewards); 3) creation of specific choice architectures (structuring environments to guide actions); 4) social influence (harnessing the power of social connections); 5) subliminal empathic messages (triggering subtle emotional responses); 6) gamification (using game-like elements to encourage habituation).

The functioning of surveillance capitalism aligns with the vision of B.F. Skinner, who, in *Beyond Freedom and Dignity* (1971), described a future centred on behavioural control while challenging principles of self-determination and individual freedom. Web giants compete for users’ attention within a business model that thrives on fostering dependency,

encouraging individuals to spend hours each day on their screens. This is achieved through behavioural techniques that stimulate dopamine release, rewarding engagement and ensuring a continuous cycle: users provide personal data, and their behaviours – both online and offline – are increasingly shaped by the platform’s suggestions.

At the heart of this system lies the activation of the reward circuitry, reinforced above all by gamification techniques designed to encourage habituation. By tapping into humans’ innate propensity for play, platforms create environments where users exchange personal information for rewards such as rankings, points, or social recognition: data that algorithms catalog to optimize targeted advertising. This dynamic resembles the logic of nudge theory, which also uses subtle suggestions and indirect assistance to guide choices. Gamification exploits the lure of immediate gratification: the pursuit of recognition, acceptance, and admiration sustains platforms such as Facebook and Instagram, where likes serve as rewards (Zuboff, 2019b, 309). Conversely, the absence of likes can prompt users to adjust strategies to gain approval and followers, thereby increasing time online. Platforms also exploit the “Fear of Missing Out” (FOMO), a form of social anxiety rooted in the fear of being excluded from rewarding experiences, which drives obsessive monitoring of social networks. Subliminal empathic messages, designed to induce targeted emotional states, further shape user behaviour. Many of these persuasion techniques (Fogg, 2003) are taught in B.J. Fogg’s “Persuasive Tech” course at Stanford, attended by designers later employed by Silicon Valley firms (including Tristan Harris, formerly an ethical designer at Google).

An emblematic example is the design of the “News Feed” on social networks, carefully engineered to foster habituation through constant refreshing. This mechanism mirrors the logic of slot machines, where intermittent reinforcement sustains engagement: just as gamblers anticipate the jackpot, users expect the dopamine rush from new posts or likes. Algorithms are optimized to maintain high engagement and guide individuals

toward behaviours aligned with the platform's commercial interests (Kramener and Guillory and Hancoc, 2014; Harris, 2016; Schull, 2019).

At this point, the inadequacy of current legislative instruments in curbing the power of Web giants and addressing manipulative practices becomes evident. Traditional legal frameworks increasingly struggle to orient citizen-users within the complex cognitive traps created by surveillance capitalists. The gap between the capacities of digital corporations and the regulatory tools available to safeguard autonomy continues to widen.⁶ Cases of mass manipulation, such as those revealed in the Cambridge Analytica scandal, highlight this inadequacy: legal analysis often reduces the issue to whether the user has given formal consent through a click, without addressing whether such consent is genuinely informed or whether data are intentionally used for manipulative or harmful purposes. The laws in force focus almost exclusively on consent, presuming a rational and independent user. Two critical aspects emerge: 1) legislation treats the absence of consent as the problem, regardless of the harms caused by manipulation; 2) current law presumptively equates the expression of will with informed consent, even when such consent cannot be considered genuinely informed. Behavioural sciences demonstrate that this presumption is a *fictio iuris*, one that indirectly produces harmful consequences for individuals and society. Overcoming this impasse requires further exploration and the development of new regulatory approaches.

5. The Traditional Law Under the Test of Behavioural Techniques of the Web Giants

Surveillance capitalism exploits human cognitive vulnerabilities for profit, deploying sophisticated influence strategies grounded in the behavioural sciences. It operates on the assumption that users, in their actual condition,

⁶ In this context, it is worth underlining the work of Marijn Sax, who develops a nuanced ethical-legal critique of data-driven digital environments. His research shows how practices that present themselves as empowering – such as health apps – often conceal manipulative dynamics, and how the “finders-keepers” logic of big data entrepreneurship rests on ethically unjustifiable assumptions about the appropriation of personal data (Sax, 2021).

are biased, irrational, and predominantly guided by System 1 thinking in their everyday decisions and interactions with digital platforms.

Within this framework, one proposed approach to addressing large-scale manipulation is to consider strategies analogous to those adopted by surveillance capitalists themselves. Unlike traditional legal frameworks, which rely on rational deliberation, these strategies target System 1 directly, acknowledging the structural limitations of decision-making agents in an online environment that is neither neutral nor transparent. It is unrealistic to expect individuals to scrutinize the privacy policies of every website they visit or to maintain constant vigilance in the face of continuous streams of manipulative messages on social networks. What becomes necessary, therefore, is the introduction of cognitive shortcuts that can guide individuals toward more beneficial choices within a socio-economic context organized around attention capture and behavioural influence.

Accordingly, the search for alternatives to traditional legal frameworks becomes crucial in order to regulate and sustain *orientation* (Stegmaier, 2019) under the pervasive influence of digital platforms. Conventional legal instruments have shown limited capacity to address the strategies employed by major technological actors, which makes it necessary to explore approaches better adapted to the dynamics of the digital environment while simultaneously protecting individual autonomy. One possible direction is the development of innovative legal mechanisms specifically designed for the digital era.

Within this discussion, the framework of libertarian paternalism and the techniques of nudging should be considered as potential instruments for counteracting manipulation. In such a counter-process, public intervention would not only address cognitive biases that lead to suboptimal decision-making but would also respond to the private interests that rely on such biases for economic gain.

In this perspective, public authorities may draw upon insights from behavioural sciences to support individuals in correcting errors of

deliberation and in adjusting their behaviour online. Nudging represents a governmental technique that avoids coercion: it relies on the automatic and reflexive mechanisms of cognition, while leaving individuals formally free to choose, with the aim of directing them toward outcomes that are expected to enhance their well-being. Nudges are particularly relevant in situations where individuals confront complex decisions, lack sufficient feedback, or cannot adequately process information – as is often the case in digital environments characterized by manipulation and information overload. According to Sunstein and Thaler, nudging is intended to provide public actors with practical instruments capable of improving the quality of individual decision-making and addressing recurring challenges of everyday life through carefully designed incentives and gentle prompts.

But also, the dark side of nudging needs to be explored. One of the most debated aspects of nudge theory concerns the suspicion that it conceals a manipulative intent. Across Sunstein's various definitions of nudges, a recurring emphasis is placed on freedom, specifically, the individual's ability to resist or escape the influence of subtle interventions. Yet such freedom may not always be guaranteed in every form of libertarian paternalism. Concerns are particularly acute where nudges explicitly exploit cognitive biases such as inertia or framing. Alongside these potentially manipulative nudges, Sunstein and Thaler also describe other interventions of a different nature: for example, self-control strategies that individuals can voluntarily adopt, or "cognitive boosts" that help counteract distortions. These gentle educational pushes aim at informing and empowering individuals rather than exploiting their weaknesses. In this sense, nudges can be distinguished into those that target System 1 (pure nudges) and those that engage System 2 (boosting) (Sunstein 2014). Because boosting techniques enhance cognitive capacity, they are less readily categorized as manipulative. By contrast, accusations of manipulation primarily concern nudges directed at System 1, which exploit cognitive biases without the goal of strengthening individual agency.

To advance the analysis, it is necessary to clarify what is meant by “manipulation”. In ordinary usage, manipulation carries strongly negative and morally charged connotations. It is often seen as a position between persuasion and coercion: the manipulator does not use force but relies on subtle means to pursue their own ends. The concealed nature of these methods renders them morally questionable, though not all acts of manipulation need be inherently condemnable. A widely accepted view (Noggle, 2006, 43-55; Noggle, 2018) defines manipulation as occurring when one person influences another by interfering with their rational faculties. In other words, A manipulates B when A elicits behaviour from B while preventing B from fully reflecting on the reasons for their actions.

This definition aptly describes certain nudges directed at System 1, particularly those that inhibit the activation of System 2, where rational capacities reside. Sunstein acknowledges these concerns and proposes a “gradualist” approach, distinguishing interventions at higher risk of manipulation (to be rejected) from those at lower risk (potentially acceptable). Yet the issue cannot be addressed by a gradualist perspective alone. A deeper concern arises when nudges promote an idealized notion of well-being that does not necessarily correspond to the individual’s own conception. In such cases, the risk of being subjected to subtle pressures that redirect desires without awareness becomes clear. Conversely, if nudges genuinely aim to promote individuals’ self-perceived well-being, suspicions of manipulation may diminish. The difficulty lies in identifying which conception of well-being guides a given intervention. Sunstein and Thaler propose a statistical conception based on majority preferences, but such preferences do not always align with those of the individual being nudged. This complicates the use of alignment between nudger and nudgee desires as a reliable criterion for judging manipulation case by case.

Manipulation is also closely tied to deception. Often manipulators disregard how the subject perceives the situation. Thus, nudges that use their influence in ways that instill false beliefs or distort preferences could rightly

be categorized as manipulative. The central concern is not influence itself, but whether it undermines deliberative and cognitive faculties. Whenever a nudge diminishes an individual's capacity for reflection and choice, it assumes a manipulative character. Nudges of this nature are clearly unsuitable for deployment by the State, whose role should be to protect, and ideally to enhance, the autonomy and self-determination of its citizens. A nudge becomes manipulative when it guides individuals – often without their awareness – toward actions shaped by deception, flawed beliefs, or reasons they do not recognize as their own. Several factors must therefore be weighed in determining whether a nudge should be rejected for its manipulative effects. Because individual autonomy is the value most threatened by manipulation, nudgers must demonstrate specific respect for the subject in order to preserve that autonomy. According to Sunstein, this is best achieved through “educational nudges”, which not only provide information but also strengthen cognitive capacities, thereby bolstering autonomy and attentional resources (Sunstein, 2014, 14-55; Sunstein, 2026, 121-168).

Not all nudges operate as educational nudges. Some gentle pushes do not respect individual autonomy to the same degree and are therefore manipulative. This occurs when nudges: 1) fail to allow individuals to independently form reasons for following the suggested direction; 2) exploit cognitive imperfections or distortions to draw attention to misleading information; 3) elicit emotions solely to make individuals vulnerable and predisposed to a particular choice. One way to mitigate the manipulative character of nudges is to eliminate their secrecy. As Sunstein suggests, transparency and publicity can reduce or even neutralize manipulative effects by making individuals more aware of the influence they are subjected to and the rationale behind it. Yet while lack of awareness often increases the effectiveness of nudges in a libertarian paternalist framework, it is also the feature that attracts the strongest criticisms regarding manipulation.

In summary, nudging carries an inherent risk of opacity and manipulation. For this reason, it should be employed only when “direct legislation”

(Bentham, 1996 [1789]; Bentham, 2010, [1802]) proves ineffective and, even then, it must be accompanied by educational policies that emancipate decision-makers from their cognitive limitations as much as possible. Nudges should thus be viewed as tools operating alongside legal instruments, complementing traditional legal mechanisms in contexts where classical methods reveal inefficiencies. The digital realm exemplifies such a context: platforms are designed to engage System 1 and make it extremely difficult for users to activate System 2. In this environment, only a tool that directly addresses System 1 can counteract manipulation and shield individuals from the strategies of surveillance capitalists.

The question, then, is not whether nudging is manipulative, but whether its manipulative capacity can serve a beneficial purpose: namely, to equitably counter the influence exerted by private actors who use behavioural techniques without regard for collateral consequences. Specific nudging interventions, combined with egalitarian education and citizen instruction on behavioural influence, could serve as genuine *awakening strategies* to protect users in their interaction with digital platforms. Bentham himself supported coupling indirect legislation with policies to educate citizens about behavioural influence techniques.

But the principal risk of nudging lies in infantilizing the nudgee. To prevent this, its application should be limited: first, to situations where there is no reasonable alternative; second, it should always be accompanied by policies aimed at educating and emancipating those subjected to nudges. Where individuals cannot realistically access the information necessary to make fully informed decisions, intervention becomes legitimate. Mill's well-known example of the bridge illustrates this point: if passers-by are unaware that the bridge is unsafe and there is no time to warn them, public officials are justified in stopping them. The same logic could be applied in digital contexts. Each time users are confronted with cookie banners, "Accept all cookies" or "Accept only necessary cookies", they lack the time and information to

evaluate the implications.⁷ A tool that speaks to System 1 could prompt the user to choose the more protective option. Yet website design overwhelmingly nudges users toward disadvantageous options: “Accept all cookies”, “Allow microphone use when the app is closed”, or “Allow camera access”. Rossi criticizes these practices as unethical, arguing that they drive individuals toward choices harmful to privacy and digital well-being (Rossi, Ducato, Hapio and Passera, 2019, 79-121). Rossi advocates for interventions in platform design to construct choice architectures that are neutral and ethical, thereby counteracting manipulative designs and orienting users toward decisions more advantageous to themselves and society:⁸ decisions they would make if given adequate time and information.

Alongside direct interventions in design, institutions must also implement educational campaigns, particularly within compulsory schooling, to inform students about human cognitive limitations and the ways in which private actors exploit them for economic gain. Yet beyond such public initiatives, it is evident that the work of orientation itself cannot be outsourced: individual commitment to sustaining orientation within this vast sea of digital manipulations is both fundamental and necessary, requiring a continuous exercise of attentiveness and responsibility.

⁷ Also Cofone argues that the contractual model based on individual consent fails within the information economy. Genuine consent is largely unattainable because the relationship between corporations and users is marked by profound asymmetries of knowledge, structural inequalities, and a lack of meaningful alternatives (Cofone, 2023).

⁸ It is relevant to underline that the literature is clear in making the point that, while design choices can indeed influence user behaviour, they are not normative in the strict legal sense, since they cannot be disobeyed in the same way as law can. In particular, Brownsword has examined how technology reshapes the regulatory environment, showing how “code” may function as a mode of governance alongside or even beyond law. Moreover, Hildebrandt combines law, philosophy, and computer science to analyse how AI and machine learning transform fundamental legal concepts such as personhood and responsibility. Finally, Lucy reflects on the fate of modern law, arguing that its abstract mode of judgment is increasingly challenged by technological modes of control (Brownsword and Yeung, 2008; Brownsword, 2019; Hildebrandt, 2015; Hildebrandt, 2020; Lucy, 2017; Lucy, 2020).

6. The “Art of Orienting Ourselves”⁹

To advance further, these issues must be placed within a broader conceptual horizon: that of *the philosophy of orientation* (Stegmaier, 2019, 155-175). According to Stegmaier, *orientation* is the ability to find one’s way and identify possible courses of action in complex or uncertain situations. It is a fundamental condition of life, comparable to breathing or nourishment, and not limited to human beings alone. Since every act of orientation occurs within a concrete context, it expresses the human capacity to keep pace with change – to make decisions that work for a time until new circumstances demand re-orientation. Orientation is always individual and fallible: one can never be completely certain of it. A philosophical account of orientation must therefore remain open, flexible, and attentive to the unforeseen. In this perspective, orientation is a continuous effort to find one’s way in always new situations: human life is a constant process of adjustment within a changing world. The philosophy of orientation connects with the sciences and with real fields such as business, politics, media, law, and art, to see how people and institutions actually deal with disorientation. It offers a descriptive approach that focuses on how attention, adaptability, and good judgment help us act meaningfully when things are uncertain¹⁰ (Stegmaier, 2019, XI-XIV, 1-3, 5-6, 15-23). Both individuals and institutions are engaged in this task of orientation, navigating uncertainty, seeking stability, and confronting pressures that shape, and often distort, their choices.

Stegmaier takes Kant’s *What Does It Mean to Orient Oneself in Thinking?* (1786) as a starting point for his philosophy of orientation. Kant had described orientation as the lived act of reason that enables us to find direction when knowledge is uncertain. Stegmaier extends this insight phenomenologically, transforming orientation from an activity of individual thought activated in specific situations of uncertainty into the basic structure of life itself. The

⁹ Stegmaier, 2023.

¹⁰ Stegmaier, 2019, XI-XIV, 1-3, 5-6, 15-23.

mind, in this view, is not a static entity but an ongoing process of orientation that connects the living being with its environment. Thinking, perceiving, and deciding are all ways of maintaining direction amid change. What distinguishes human consciousness is its ability to reflect on its own orientations – to orient itself within its orientation – thus making philosophy a practice of attentiveness to how we find our way in the world. In Stegmaier’s framework, a subtle but crucial distinction emerges between *Bewusstsein* (consciousness) and *Bewusstheit* (awareness). *Bewusstsein* refers to the general capacity of living beings to orient themselves in the world: the ongoing, pre-reflective process through which perception, action, and judgment are coordinated. *Bewusstheit*, by contrast, designates the variable intensity or clarity with which orientation becomes fully aware (Stegmaier, 2019, 80-81). One might cautiously compare this distinction to Kahneman’s dual-system model: *Bewusstsein* roughly corresponds to the fast, intuitive “System 1”, while *Bewusstheit* resembles the slower, reflective “System 2”, which is activated when uncertainty or surprise increases and re-orientation is required. For Stegmaier, routines are not mere repetitions but forms of practices of the self that sustain orientation in everyday life (Stegmaier, 2019, 77-82). They provide temporary stability, allowing individuals to remain attentive and to renew their sense of direction when circumstances change. In this way, routines cultivate a background that supports *Bewusstheit*, the reflective awareness that emerges when one becomes conscious of one’s own orientation. This idea recalls Foucault’s notion of “practices of the self” (Foucault, 1984; Foucault, 1988), understood as intentional exercises – such as writing, meditation, or self-examination – through which individuals work upon themselves to shape their existence. Both perspectives emphasize self-formation as an ongoing, situated activity: while Foucault highlights the historical and ethical dimensions of these practices within power relations,

Stegmaier approaches them phenomenologically, as everyday forms of maintaining awareness and self-direction amid uncertainty.¹¹

Such an individual and reflective approach may constitute a crucial dimension of resistance to the contemporary forms of behavioural manipulation enacted by the “surveillance capitalists”. The deliberate cultivation of awareness and the disciplined maintenance of routine – conceived also, for example, in analogy to mechanisms of “pre-commitment”¹² – can function as a personal and ethical countermeasure to external architectures of influence. These practices may complement collective and institutional interventions by fostering an inner capacity to remain oriented and attentive amid the continuous and often insidious cognitive traps that characterize today’s informational environments.

This art of orientation can play a crucial role in enabling citizen-users to awaken their attention and become more self-aware of their cognitive vulnerabilities in the digital age, which has become an era of disorientation, marked by phenomena such as:

the collection and storage of vast amounts of data regarding every internet user that he or she discloses in some way or that can be

¹¹ “The evolution of orientation proceeds via the succession of self-stabilization, destabilization, and re-stabilization: routines and routine patterns develop, fail, and develop in new ways. If the re-stabilization of routines or of whole orientation worlds (like politics, sports, or family) fails, they can be dismissed. Not only footholds and routines, but also orientation worlds are selected at the cost of some and the benefit of others (preferring one’s family life over sports, sports over politics, or politics over one’s family life)” (Stegmaier, 2019, 91). “Self-bindings may also be of a non-moral kind; habits or routines are also permanent self-bindings that exclude other options” (Stegmaier, 2019, 208).

¹² The concept of pre-commitment refers to a strategy by which an individual voluntarily restricts their future freedom of choice to avoid acting against their long-term interests — for instance, yielding to temptation, procrastination, or cognitive bias. First theorized by Thomas Schelling (1960), it describes how people or institutions can bind themselves in advance to make their future behaviour more consistent or credible, like Ulysses tying himself to the mast to resist the Sirens. In behavioural economics, pre-commitment devices (such as savings plans or digital self-control tools) help counter *time inconsistency*. In law and political theory, constitutions or regulations can be seen as collective pre-commitments, limiting future decisions to preserve rationality or stability (Schelling, 1960; Elster, 1979).

gathered from his or her online behaviour and then used to influence his or her buying or voting behaviour (2019, 256).

According to Stegmaier,

Big data can reorient our interindividual, doubly contingent orientation to other orientations to a unified macrosocial orientation that is governed by specific organizations (ibidem).

Altogether, the digitization of human *Orientierung*¹³ creates a fundamentally new situation of orientation, one in which the very conditions of finding and maintaining direction are being reshaped. Digital technologies expand the potentials of orientation by granting access to vast amounts of information, instantaneous communication and algorithmically tailored and personalized guidance systems. At the same time, they exponentially increase the needs for orientation, since the sheer abundance of data, the acceleration of communication and the opacity of algorithmic structures overwhelm traditional cognitive and social capacities for making sense of the world. Never before have the opportunities for orientation and the risks of disorientation been so closely intertwined. This situation places unprecedented demands on our abilities of orientation, requiring not only technical skills and digital literacy but also new forms of ethical reflection, critical judgment, and philosophical guidance. Digital technologies externalize the rationalization of unconscious drives to third parties, so that we end up acting not on impulses we have individually processed, but on impulses that are scientifically directed toward the fulfillment of others' interests. This displacement of self-regulation transforms the very structure of agency: awareness and decision-making are no longer grounded in individual reflection but mediated through algorithmic architectures that anticipate, predict, and channel behaviour. In this context, regaining our capacity for orientation means developing everyday *practices of the self* that

¹³ German original word for "Orientation" (Stegmaier, 2008; Stegmaier, 2019).

help us keep a reflective distance and act with greater awareness. In short, digitization magnifies both our capacity and our vulnerability, making the task of orientation more complex, more urgent, and more indispensable than ever (Stegmaier, 2019, 258; Stegmaier, 2023; Mueller, 2023):

The so-called internet of things, where machines autonomously communicate with each other in order to optimize all kinds of processes, will, as we can already predict, go far beyond this and influence human life on a much deeper level. In the long run, the evolution of humanity then may turn into an independent evolution of machines. The fears are not unjustified; in them, an old fear of humankind returns on a grander scale: that a situation will be created that humankind can no longer master and at whose mercy it will then be entirely powerless. It is the fear of the end of one's own orientation.

To survive in this era, everyone must discover an individual way to orient themselves – an art of living and navigating in a complicated world filled with compelling traps for our weak and paleolithic brains. Institutional action is necessary, yet it is not sufficient. In a world where every certainty collapses and the law appears outdated, such measures alone cannot protect people from the vast and massive manipulation of minds. Each individual must cultivate his or her own way to remain focused and to resist being disoriented by the manifold tools of influence. This art of orientation should be an art of living, cultivated as a daily habit. It also requires knowing how our brains function – where they are strong and where they are weak – and learning how to rely on System 2 rather than System 1. Through small and precise practices, we can resist the pervasive power of influence and maintain our capacity for self-determination.

According to Stegmaier, power becomes most visible in those moments when we are unable to master a situation on our own, when circumstances overwhelm us and reveal the limits of our control. While individuals can only

partially overcome entrenched social and cultural power relations, they are not entirely without resources. Even within such constraints, it is often possible to carve out new spaces of freedom, to open leeways in which alternative courses of action can be pursued. In doing so, individuals may gradually weaken or destabilize existing structures of power, subverting them from within and preparing the ground for more deliberate forms of resistance or even collective revolts.

In this sense, the art of living does not consist in the illusion of total mastery but in the subtle practice of orientation: learning how to navigate pressures and asymmetries, how to discern possibilities in restrictive environments, and how to transform limitations into openings for action. Orientation thus becomes a strategic art, one that allows individuals to endure, adapt, and, when possible, redirect the very forces that seek to dominate them.¹⁴

In conclusion, only a symbiotic relationship between institutional intervention and individual awareness can serve the task of awakening minds and safeguarding them against disorientation and the pervasive influence exerted by surveillance capitalists. This relationship can be described and summed in these points: 1) nudging as a technical tool: when used judiciously, nudging can complement traditional legal systems. It is designed to speak directly to the fast, intuitive processes of System 1, especially in situations where individuals lack the time or information to make fully informed decisions. Nudges can counteract the influence of powerful digital actors and guide users toward choices better aligned with their preferences, privacy, and well-being; 2) educational policies: educational measures are pivotal in raising awareness of cognitive limitations and vulnerabilities. By exposing the psychological traps set by digital platforms, such policies

¹⁴ “Hence, you must ultimately orient yourself on your own. Orientation – as the achievement of finding your way in a new situation in order to find possibilities for actions through which you can master the situation – is the achievement of an individual ability; and with your individual ability, you may cope with one situation but not with another. Every orientation may face unforeseen and surprising circumstances, where its abilities might fail. Therefore, one can never be entirely sure of one’s orientation” (Stegmaier, 2019, 6-7).

empower users to make more informed decisions. Compulsory schooling, in particular, offers an ideal venue to equip students with the knowledge and critical thinking skills required to navigate the digital world; 3) philosophical reflection: philosophy – especially the philosophy of orientation – can provide individuals with a deeper understanding of the ethical and moral dimensions of digital interactions. Encouraging critical thinking enables individuals to reflect on the values that underlie their choices and to assess their broader implications. Such reflection can guide users in making decisions consistent with their ethical principles; 4) self-awareness of cognitive vulnerabilities: self-awareness is essential for cultivating a keen sense of one's cognitive weaknesses. Recognizing these vulnerabilities, and understanding how they are exploited by digital platforms, allows users to become more resilient and resistant to manipulation; 5) navigating the digital landscape: through this holistic approach, individuals are better equipped to discern the subtle techniques employed by digital corporations and to make choices that safeguard their well-being and autonomy.

To sum up, on the collective and institutional level, nudging could operate as a form of *counter-nudge*: public authorities might deploy behavioural insights not to encourage consumption or maximize engagement, but rather to support autonomy, well-being, and informed decision-making. And on the other hand, on the individual level, the philosophy of orientation offers a framework for cultivating self-awareness and decision-making skills, enabling people to recognize manipulative techniques and maintain control over their choices.

While the philosophy of orientation offers a normative-ethical guide for judgment, nudging provides a repertoire of practical interventions that support reflective decision-making without undermining freedom. Taken together, these elements promote a conception of agency that is both context-sensitive and resilient – particularly vital in the digital domain, where individuals face growing exposure to algorithmic influences and epistemic asymmetries. In this context it is worth emphasizing that nudging is a social

technique that can be used both to orient and to disorient. Some nudges act directly on “System 1” to steer behaviour automatically, while others are designed to awaken “System 2” and foster reflective awareness. Developing appropriate and ethically sound forms of nudging therefore requires clarifying what it means for individuals to regain awareness of their own decision-making processes – and for this purpose, the philosophy of orientation provides a valuable conceptual framework.

The aim of this paper has been to bring these questions into philosophical focus and to sketch possible trajectories for addressing them. Although preliminary in scope and leaving several lines of inquiry open, it is conceived as a conceptual starting point for a broader and more sustained research agenda. Future work could further develop the phenomenological implications of Stegmaier’s philosophy of orientation in relation to the challenges of digital environments, behavioural governance, and surveillance capitalism. It might also investigate how individual practices of awareness and routine – understood as contemporary forms of “practices of the self” – can interact with collective and institutional frameworks to counteract cognitive manipulation and sustain autonomy. In this sense, the present contribution aspires not merely to describe a theoretical constellation but to outline a path for continued interdisciplinary reflection on how to remain oriented within increasingly complex architectures of influence.

References

- Bentham J. 1996 [1789]. *An Introduction to the Principles of Morals and Legislation*. Edited by J. H. Burns and H. L. A. Hart (Clarendon Press).
- Bentham J. 2010 [1802]. *Traité de législation civile et pénale* (Dalloz).
- Brownsword R. (2019). *Law, Technology and Society: Re-Imagining the Regulatory Environment* (Routledge).
- Brownsword R. and Yeung K. (eds.) (2008). *Regulating Technologies: Legal Futures, Regulatory Frames and Technological Fixes* (Hart Publishing).

- Byung-Chul H. (2017). *Psychopolitics: Neoliberalism and New Technologies of Power* (Verso).
- Cofone I. (2023). *The Privacy Fallacy: Harm and Power in the Information Economy* (Cambridge University Press).
- Cohen J. E. (2019). *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press).
- Couldry N. and Mejias U. A. (2019). *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism* (Stanford University Press).
- Doctorow C. (2021). *How to Destroy Surveillance Capitalism* (Verso).
- Elster J. (1979). *Ulysses and the Sirens: Studies in Rationality and Irrationality* (Cambridge University Press).
- Fogg B. J. (2003). *Persuasive Technology: Using Computers to Change What We Think and Do* (Morgan Kaufmann).
- Foucault M. (1988). *Technologies of the Self* (University of Massachusetts Press).
- Foucault M. (1984). *The Care of the Self* (Gallimard).
- Gödde G., Zirfas J., Mueller R. G. and Stegmaier W. (eds.) (2023). *Nietzsche on the Art of Living: New Studies from the German-Speaking Nietzsche Research* (Orientations Press).
- Harris T. (2016). How Technology Is Hijacking Your Mind – from a Magician and Google Design Ethicist, in *Medium*.
- Hildebrandt M. (2020). *Law for Computer Scientists and Other Folk* (Oxford University Press).
- Hildebrandt M. (2015). *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Edward Elgar).
- Kahneman D. (2011). *Thinking, Fast and Slow* (Farrar, Straus and Giroux).
- Kahneman D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics, in *The American Economic Review*, vol. 93(5), 1449–1475.

- Kahneman D. and Tversky A. (eds.) (2000). *Choices, Values, and Frames* (Cambridge University Press).
- Kahneman D. and Tversky A. (1981). The Framing of Decisions and the Psychology of Choice, in *Science*, vol. 211(4481), 453–458.
- Kramener A. D. I., Guillory J. E. and Hancock J. T. (2014). *Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks* (Princeton University Press).
- Lucy W. (2022). The Death of Law, Another Obituary, in *Cambridge Law Journal*, vol. 81, 350.
- Lucy W. (2017). *Law's Judgement* (Hart Publishing).
- Middleton R. (2016). Pokemon Go: Dutch Rail Operator Tells Nintendo to Change Game after Players Wander onto Tracks, in *International Business Times UK*, 12 July.
- Minda G. (1994). *Postmodern Legal Movements: Law and Jurisprudence at Century's End* (New York University Press).
- Morozov E. (2019). Capitalism's New Clothes: Shoshana Zuboff's Surveillance Capitalism, *The Baffler*, 4 February.
- Mueller R. G. (2023). Nietzsche Art of Living in the United States Today, in G. Gödde, J. Zirfas, R. G. Mueller and W. Stegmaier (eds.), *Nietzsche on the Art of Living: New Studies from the German-Speaking Nietzsche Research* (Orientations Press).
- Noggle R. (2018). The Ethics of Manipulation, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.
- Noggle R. (2006). Manipulative Actions: A Conceptual and Moral Analysis, in *American Philosophical Quarterly*, vol. 33(1), 43–55.
- Revesz R. (2016). Pokémon Go: Gamers Warned to Pay Attention to the Law When Searching for 'Pokéstops', *The Independent*, 12 July.
- Rossi A., Ducato R., Hapio H. and Passera S. (2019). When Design Met Law: Design Patterns for Information Transparency, in *Droit de la Consommation*, vol. 1, 79–121.

- Sax M. (2021). *Between Empowerment and Manipulation: The Ethics and Regulation of For-Profit Health Apps* (Wolters Kluwer).
- Schelling T. C. (1960). *The Strategy of Conflict* (Harvard University Press).
- Schüll N. D. (2019). Your Undivided Attention, *Podcast Episode 1: What Happened in Vegas*, Center for Humane Technology.
- Simon H. A. (1982). *Models of Bounded Rationality* (MIT Press).
- Simon H. A. (1955). A Behavioral Model of Rational Choice, in *Quarterly Journal of Economics*, vol. 69, 99–188.
- Stegmaier W. (2023). The Art of Living as an Art of Orientation, in G. Gödde, J. Zirfas, R. G. Mueller and W. Stegmaier (eds.), *Nietzsche on the Art of Living: New Studies from the German-Speaking Nietzsche Research* (Orientations Press).
- Stegmaier W. (2019). *What Is Orientation? A Philosophical Investigation* (de Gruyter).
- Stegmaier W. (2008). *Philosophie der Orientierung* (de Gruyter).
- Sunstein C. R. (2016). People Prefer System 2 Nudges (Kind of), in *Duke Law Journal*, vol. 66(1), 121–168.
- Sunstein C. R. (2014). *Why Nudge? The Politics of Libertarian Paternalism* (Yale University Press).
- Sunstein C. R. (2014). *The Ethics of Nudging*, Harvard Public Law Working Paper No. 14-55.
- Sunstein C. R. (2013). *Simpler: The Future of Government* (Simon & Schuster).
- Sunstein C. R. (2013). The Storrs Lectures: Behavioural Economics and Paternalism, in *The Yale Law Journal*, vol. 122(7), 1826–1879.
- Sunstein C. R. and Thaler R. H. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale University Press).
- Sunstein C. R. and Thaler R. H. (2003). Libertarian Paternalism Is Not an Oxymoron, in *University of Chicago Law Review*, vol. 70(4), 1159–1202.
- Wilson E. O. (2009). Debate at the Harvard Museum of Natural History, 9 September 2009.

Zuboff S. (2019a). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs).

Zuboff S. (2019b). *The Age of Surveillance Capitalism* (Profile Books).

Zuboff S. (2014). A Digital Declaration, in *Frankfurter Allgemeine Zeitung*, 5 March 2014.

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

Paternalistic Interventions: What do They Presuppose About Human Rationality, and When are They Justified?

MIGUEL FERNÁNDEZ NÚÑEZ

Professor of Philosophy of Law, Autonomous University of Madrid (Spain)

✉ miguel.fernandezn@uam.es

ORCID <https://orcid.org/0000-0001-7870-1817>

ABSTRACT

In this article, I examine a pressing and recurring problem for liberal thought: paternalistic interference with freedom. I focus on its main premise: a deficit in the affected agent's deliberation that results in harm to the agent. I analyse how human beings are viewed by liberal legal norms — views about the effective rationality of normative subjects. I identify and characterise the four necessary and sufficient conditions to justify a paternalistic intervention. I outline several scenarios involving different defects in individual deliberation by identifying the agent's true self and basic interests. Finally, I contrast this repertoire with projects to build an “anthropology in law”, such as the one recently proposed by Moreso. My critical evaluation concludes that attributing “vulnerability” generically or hastily, as Moreso does, to the human collective, without precisely determining deliberative flaws, can lead to an excessive expansion of paternalistic interventions and an undue restriction of individual autonomy.

Keywords: paternalism, political liberalism, justifications of legal norms, legal presuppositions, rationality deficits, José Juan Moreso

This article is the result of a research stay at the Faculty of Law, University Roma Tre, where a previous version was discussed on 21st May 2025. This paper was written within the framework of the research project EcoGentium. Ecological justice and new vulnerabilities: global legal challenges (2024/00209/001), funded by the Agencia Estatal de Investigación (AEI). I would like to thank Juan Carlos Bayón, Ezequiel Monti, José Juan Moreso, Giorgio Pino, Michele Ubertaino and especially Giuseppe Rocchè and an anonymous reviewer from Athena for their comments on earlier drafts of this paper. Their suggestions were essential in enhancing the accuracy and depth of my article.

ATHENA

Volume 5.2/2025, pp. 142-185

Articles

ISSN 2724-6299 (Online)

<https://doi.org/10.60923/issn.2724-6299/22521>



“Paternalistic intervention must be justified by the evident failure or absence of reason and will; and it must be guided by the principles of justice and what is known about the subject’s more permanent aims and preferences, or by the account of primary goods. [...] Paternalistic principles are a protection against our own irrationality, and must not be interpreted to license assaults on one’s convictions and character by any means so long as these offer the prospect of securing consent later on.”
J. Rawls

1. Introduction: The Problem of Paternalism for Classical Liberalism

1.1 Paternalism. What is it?

We can call “paternalistic” the intervention through which an agent (typically the state) interferes with another agent’s liberty, justifying the interference as being in the latter’s interest.¹

Thus, they are the justifications (the intentions, the aims) behind the decisions that are paternalistic; such a decision may take the form of a moral or legal norm, a political or administrative measure.²

The standard normative effects of paternalistic norms and measures are obligations and prohibitions. However, there are also good reasons to include normative disabilities. For instance, the state’s refusal to confer legal validity on certain contracts, such as a contract of voluntary slavery.

A given justification is paternalistic if and only if it derives from the paternalistic principle.

¹ As long as an agent (hereafter, the state) interferes with an individual’s freedom of action, such interference is to be considered coercive. To be sure, this is “coercion” in a minimal sense. I will discuss these topics shortly.

² Therefore, it is clear that expressions such as “paternalistic interventions (decisions, norms)”, which are common in specialised literature, are metonyms (see Diciotti, 2021, 97; cf. Diciotti, 1988, 85). Nevertheless, for the sake of brevity, I often use them in this paper. Setting aside the metonym, a decision can be called “paternalistic” if and only if its sole or main justification is paternalistic.

Paternalistic principle³ establishes that if an individual *X*'s action harms her interests, the agent *Y* (the state) should *pro tanto* intervene.⁴

1.2 Justifications of Decisions in the Context of Liberalism

As stated, in this article I analyse the paternalistic principle within the normative framework of classical liberalism. By “classical liberalism” – or simply “liberalism” from now on – I refer to the ethical-political doctrine paradigmatically articulated by Mill, one of whose fundamental theses is precisely individual liberty and the principle of autonomy.⁵

As is well known, one of the main tenets of liberalism is that the state should guarantee individuals a domain of rights against invasion by others, and Mill's starting point is the defence of the individuals' “sovereignty” over their personal sphere, except for those actions that cause harm to third parties (Mill, 1864, 25). Nonetheless, Mill soon finds himself qualifying his initial theses towards paternalism, which itself highlights the problem we are concerned with (notably Mill, 1864, 172 ff.).

Very broadly, the ideal of autonomy — the idea that the individual should become “the author of her own life”, reflexively and free from significant external interference (primarily and above all from state coercion) — is the foundation of liberalism, as is individual liberty, insofar as it is instrumentally effective in achieving autonomy. This implies that liberalism requires specially qualified justification when it aims to restrict liberty (Moreso, 2021a, 650).

³ Such a principle is, of course, ethico-political; therefore, moral norms are its paradigmatic embodiment. In this article, I do not specifically refer to “legal paternalism” (i.e., the paternalistic justification of legal norms) or to “moral paternalism”, assuming that the differences between them are secondary. Nevertheless, I will occasionally point out some differences when they are relevant.

⁴ “Should intervene” means that such action is normatively required. Alternatively, we can phrase it as “the state has reasons for acting” or “the state is obliged to act”, and so on. The theoretical framework applied to this principle is not important to our concerns.

⁵ I leave aside the fact that classic liberalism is a set of minimally heterogeneous proposals. This is an issue of the reconstruction of the history of ideas and can be ignored for the sake of expository economy. However, there is one point theoretically relevant to this paper: the contradiction between some theses put forward by Mill and Kant, forefathers of liberalism. I will address this issue specifically at the end of my essay (§3.4).

Indeed, liberalism adopts different views of a given intervention depending on its justification (see Feinberg, 1984, 26-27, for a similar presentation):

1. Harm principle establishes that, when the action of an individual (say Giuliana) harms another (say Gabriela), the state should *pro tanto* interfere with Giuliana's action. In other, more explicit words, if Giuliana causes or intends to cause harm to Gabriela and it is not permissible for Gabriela to suffer that harm, either if Giuliana's rights do not cover the harmful action or the normative considerations (whether substantive or procedural) in favour of preventive or reparatory interference outweigh those against it, then the state is all things considered justified in interfering with Giuliana's action.
2. Paternalistic and anti-paternalistic doctrines are concerned with the scenario in which Giuliana's action causes her harm. However, a distinction should be made based on Giuliana's rational capacities:
 - 2.1. Anti-paternalism establishes that, if she can discern, then *ceteris paribus* the state should not interfere with her action.
 - 2.2. Soft paternalism, instead, establishes that, if she is unable to discern, then *ceteris paribus* the state should interfere with her action.⁶

This holds in most cases (hence the *ceteris paribus* clause), but there can be exceptions.⁷

- 2.3. Genuine (hard or moderate) paternalism seeks to answer two questions: If she is formally and materially (at least normally) able to discern, but finds herself in a situation of (temporary or partial) material

⁶An exemplar of this is belonging to the category of *alieni iuris* subjects. The inclusion of a subject in a category of people who are formally incapable establishes the presumption that the subject is also materially incapable. This is a normative qualification which is often *iuris et de iure*, i.e., treated as indefeasible. Such a presumption is plausible as far as it is based on statistical generalisations, correlating the occurrence of certain empirical situations with a sufficient lack of relevant cognitive or practical abilities.

⁷ Of course, it should be recalled that the question of formal (in)capacity is clear-cut – and, in addition, it must be so, as required by the material dimension of the legality principle – but this is not the case for material (in)capacity, as there is a continuum from radical incapacity to total capacity.

incapacity, should the state interfere with her action? Under what conditions?

As we will soon find out, [2.3] can be regarded as specifying the *ceteris paribus* clause in [2.1]: what changes when other things are not equal, namely, the conditions of rationality.

3. Anti-moralism establishes that, if Giuliana performs an immoral action, the state *pro tanto* should not interfere with her action. In other words, its immoral content is not in itself a basis for the state to intervene in a given action.

4. Anti-perfectionism establishes that, if Giuliana performs an action that contradicts a principle shaping the idea of “how Giuliana should become” – an idea not shared by Giuliana herself –, the state *pro tanto* should not interfere with her action.⁸

With the exception of [1], in which I specified the (usual) conditions under which the intervention measure should be taken all things considered, the remaining justificatory principles are merely *pro tanto*. Certain elements within the aforementioned definitions remain implicit and should therefore be spelled out. Furthermore, for these to become all things considered principles, the procedural and substantive considerations against state intervention must not outweigh or otherwise counterbalance the considerations in favour of state intervention. I will return to those that I call “background conditions”.

1.3 Where the Problem Lies: The Opportunity to Decide for Others, Based on Incapacity and Hypothetical Consent

⁸ As with paternalism, other forms of interaction between the state and citizens are certainly conceivable and feasible. The state may wish to persuade citizens – in a discursive, rhetorical, or dialogical sense, and therefore without resorting to coercion – to make certain choices so that they become the citizens they “ought to be” (whatever this may mean, which is not important here). Some persuasive strategies or some ways of exercising regulatory powers for the benefit of the competent agent herself (in a paternalistic or perfectionist sense) that are less forceful than coercion can be legitimate. However, this is a hypothesis I do not intend to explore in this work, except occasionally.

Let us return to the arguments listed above. Several observations should be made. Preliminarily, it should be emphasised that justifications can be made explicit, but they can also be implicit. Even when explicit, their formulation can be sufficiently indeterminate to allow for cumulative and alternative justifications, not to mention the possibility of manipulation by the interpreter. It is very easy for the same decision to be presented, either by the authority that promulgates it, that called upon to interpret it, or for the observer or critic, to identify, alongside paternalistic foundations, non-paternalistic foundations that are (or that are intended to be) complementary or alternative to the former.

In the second part of this article, through the critical analysis of the Lochner and Wackenheim cases, it will be sufficiently illustrated how one and the same judicial decision can be seen as conveying different justifications. In particular, we will see how the conceptual boundaries of paternalism with the other doctrines (namely moralism and perfectionism) are blurred; however, it is advisable to consider these definitional boundaries from the outset, as well as the autonomous conceptual profile that paternalism appears to have.

It is certainly difficult to distinguish perfectionism from unjustified paternalism, insofar as both apply standards to the subject that are foreign to her (cf. Alemany, 2006, 114-115, 162-163). On the other hand, perfectionism differs from justified paternalism because, as I will argue, only for the latter position, is “the view Giuliana has of herself” or “the view Giuliana would like to have of herself” important (cf. Hart, 1963, 30-31; Nino, 1989, 414, 427, 430-431; Alemany, 2006, 384-386). In contrast, [3] and [4] are positions that, *ex hypothesi*, exclude taking into consideration the agent’s motivational set.

With regard to [2], i.e. the paternalistic scenarios, in this work I focus on [2.3], which is the position that is most strongly or, more precisely, genuinely problematic for liberalism. [2.3] embodies the conceptual space occupied by what are generally called “hard paternalism”, “moderate paternalism”, or

“moderate anti-paternalism” (cf. Maniaci, 2025, 14-15). No doubt, we must determine to what extent the state should intervene, and thus which position regarding intervention is correct.

As for [2.1], as we shall see in a moment, it is a clear case of anti-paternalism for the liberal, i.e. a case where paternalism is not justified for the liberal. Regarding [2.2], the literature usually refers to such cases as “weak” or “soft paternalism” (Feinberg, 1986, 12 ff.; Alemany, 2006, 392 ff.). In a sense, it is a case where paternalistic intervention is clearly justified for the liberal. Indeed, in some of these scenarios, if we focus only on the limiting case where the individual is not autonomous at all, there are arguments to suggest that it is wrong to speak of “paternalism” if there is no autonomous exercise of liberty to interfere with. Given that, for the potential beneficiaries of soft paternalism, the assertion of a deficit of rationality and therefore the need to interfere with an embryonic, weak autonomy is more indisputable, the considerations applying to strong paternalism, i.e. the most problematic type of paternalism, will apply even more so (“*a fortiori, a maiori ad minus*”) to weak paternalism.

To better understand why [2.3] is the truly interesting position, it is worth examining the assessments in [2] in greater depth and considering their merits. Indeed, there is a connection between the incapacity of the subject and the consent that the subject can give or, ultimately, that can be considered valid. The underlying idea is that if we want to count a certain exercise of freedom as autonomous, it must satisfy the appropriate conditions (cf. Elster, 1983, 20 ff.). In [2.1], it is implied that Giuliana consents to the action she carries out; in [2.2], it is implied that Giuliana, given her incapacity, would not consent to her action if she were to abandon her state of incapacity.

The question in [2.3] is more difficult for two reasons. Firstly, unlike in [2.2], it is arduous to argue that the individual finds herself in a state of incapacity. Although she could be so, the incapacity must be qualified: what type it is, and how pervasive it is. For brevity, we can refer to this as the “problem of diagnosis”.

Nevertheless, the result of [2.3] aligns with that of [2.2]: *justified* cases of paternalistic intervention always presuppose some form of *rationality* deficit.

Indeed, the two questions that form the subtitle of this article are examined thoroughly, albeit far from exhaustively, and combined: the justification for paternalistic interference therefore always rests on some flaw in the agent's deliberation, along with other related, secondary conditions. Of course, it is important to note that not every deliberative flaw is sufficiently damaging; in a second step, it should be examined to what extent the flaw or the combination of flaws is relevant to justify paternalistic intervention. I will offer some indications regarding this question, especially concerning the interests affected by these flaws.

Incidentally, two presuppositions of paternalistic intervention must be explicated here: that the state knows better than the agent what is best for her well-being,⁹ and that the state is precisely the one called to intervene.¹⁰

Concerning the potential alignment between [2.1] and [2.3], a further consideration is necessary. As argued in [2.1], it is presumed that Giuliana consents to the action she undertakes because, as stated, "Giuliana is able to discern". Nonetheless, it should be noted that the fulfilment of specific criteria is requisite in order to proceed. As Maniaci observes, moderate anti-paternalists allow individuals to freely pursue their well-being as they deem appropriate, provided they are

sufficiently rational [mainly, but not only in the sense of "coherent"], aware of the relevant facts, and sufficiently free from coercive pressures [and compulsion], and it must be the case that the desires

⁹ As opposed to the classic "argument from paternalistic distance" (cf. Diciotti, 1986, 566). This claim will often prove false, especially since the state (e.g., the legislator) relies on empirical generalisations – exposed to exceptions and sometimes even refutable – and applies them to general and abstract categories of normative subjects. This caveat is of little importance: it is an easily granted and useful assumption from an explanatory-theoretical perspective for the purposes of normative arguments such as those examined and put forward in this paper.

¹⁰ Indeed, while we can and should expect paternalistic decisions from a parent, the issue is whether the state should act as a "*pater*" towards its citizens.

they form are stable over time (see Maniaci, 2025, 5; cf. Maniaci, 2025, 13-15).¹¹

Regarding the adjective, such an anti-paternalist is rightly labelled “moderate” because she recognises that certain conditions must be satisfied to respect the individual’s decision. Regarding the substantive, the anti-paternalist (even if moderate) remains anti-paternalist insofar as she presumes the rationality of agents and reviews it only when there is evidence of a potential deficit of rationality. I believe it is correct to assume the general rationality of subjects that belong to the class of [2.1]; nonetheless, in some problematic cases, we set aside this presumption and must carefully examine their competence. In those exceptional cases, we are already entering [2.3].¹² This does not deny the utmost importance, for liberals, of assuming individuals’ *pro tanto* material capacity in performing any action affecting only themselves. For the purpose of [2.1], *pro tanto* respect toward individuals’ self-referring decisions is required. Moreover, as I will show in this article, other conditions must be met to justify all things considered paternalistic interventions.

Finally, while in [2.2] it is undisputed that such consent is invalid – what is problematic, if anything, is whether the subject reaches a formal threshold –, it is debatable whether, in the absence of such incapacity (which is at issue in [2.3]) the subject would unfailingly consent to the intervention. The core task is to determine under which conditions the agent’s explicit consent to a supposedly free course of action *X* can be considered invalid or even non-existent, and under which conditions the subject would consent to the

¹¹ For a detailed disclosure of such exigences, see Maniaci (2025, 89-106).

¹² A potential distinction between Maniaci’s perspective – and other analogous perspectives, such as Alemany’s (2006, 422-423) – and my own pertains to the fluidity of the scenario delineated between [2.1] and [2.3]. This is due to the fact that, as Maniaci would acknowledge, it is possible to make deliberative errors of a significant nature without being formally incompetent, i.e. [2.2] (cf. Maniaci, 2025, 94-95). While my position may appear to adopt a more paternalistic stance than that of Maniaci when considered within the confines of the initial conditions for intervention, this is not the case when the procedure is continued to its culminating conditions and the various caveats attached to each condition are taken into account.

paternalistically interfered course of action *Y*. There is a phenomenology of deliberative elements involved, as well as a series of symptoms, that lead us to identify this consent (primarily, the basic interests). We can refer to this as the “problem of hypothetical consent”.

1.4 Overview of the Paper: The Four Necessary and Sufficient Conditions for Paternalistic Intervention

In the first part of this paper, I intend to address the problem of decision-making on behalf of others (hereafter referred to as “hetero-decision”) and, particularly, the problem of diagnosis and the problem of hypothetical consent. After examining the definitional questions above, I will focus on the problem that hetero-decision in the agent’s interest poses for liberalism. Next, I will show that the first step in developing and analysing a procedure for paternalistic intervention is to balance well-being and autonomy correctly. For paternalistic interference with freedom to be justified, well-being must outweigh or otherwise counterbalance autonomy. I will then discuss the connection between the subject’s well-being and the need for paternalistic intervention, developing a brief phenomenology of flaws in deliberation and the most important types of interests that can be affected by them.

In other words, I will present a brief case study of situations in which an individual is incapacitated, considering the more and less problematic aspects of paternalistic intervention in these situations. Only when the agent incurs some rationality deficit can intervention with her freedom of action be justified. This is the second step, or necessary condition, of paternalistic intervention.

In particular, I will explore a way to articulate the conceptual link between the subject’s well-being and the consent they would give to paternalistic intervention. Determining the affected agent’s “true self” and identifying it with the agent’s basic interests constitutes the third step or necessary condition for paternalistic intervention.

The fourth step consists in satisfying the “background conditions”, that is the substantive and procedural factors in favour of intervening itself should outweigh the opposing considerations. I will set out several considerations that give shape to these background conditions.

In summary, I aim to show that a paternalistic intervention is justified if and only if: [1] it results from correctly weighing well-being against autonomy (that is, if well-being prevails, while autonomy is rightly outweighed or otherwise counterbalanced, or not ultimately sacrificed, even if it appeared so *prima facie*); [2] it arises from one or more scenarios of rational deficits in a sufficiently damaging way; [3] it is consistent with the agent’s motivational set, particularly, in her “true self” and basic interests; [4] the procedural and substantive reasons for intervening outweigh those against intervening. Thus, these four steps, or sets of criteria, are what a procedure of paternalistic intervention must satisfy to be justified. I will address [1], [2] and [3] at some length, though not exhaustively or with the ideal depth, and I will provide several indications to determine [4].

In the second part of the essay, I will examine the attempt to build an anthropology in law. By “anthropology in law”, I refer to the assumptions of rationality in legal norms, and the explicit or implicit representations in norms of human beings as more or less rational agents.¹³ I will also examine how certain proposals seek to defend paternalism on the basis of a supposedly realistic (true or plausible) anthropology that the law must consider. I will focus on the recent proposal by José Juan Moreso, by pointing out how Moreso’s proposition is affected by several problems.

In general, and preliminarily, the very project of constructing a single anthropology in law can be considered questionable and, indeed, misleading. Instead, there is a theoretical need to clearly define and list cases of rational vulnerability; there is also a theoretical and practical need to examine

¹³ These may be the (primary or secondary) addressees of the legal norms. The status of these statements which make up anthropology in law appears to be hybrid: in part, they are apophantic statements, concerning how normative subjects actually are; in part, they are evaluative statements, concerning what treatment should receive them.

remedies for such cases of vulnerability in a singularised way. As they are descriptively inadequate, and as they are theoretically and practically counterproductive, it will be necessary to reject those overly general anthropological accounts based exclusively on the image of a fully capable subject – such as defending the extrapolation to any area of law of the assumption of a *sui iuris* subject with full capacity to act, as a contractual party under the Civil Code –, and those overly general anthropological accounts based on the image of a (generically) incapable, fragile subject. Moreover, as we shall see, Moreso’s conclusion is based on a generic reconstruction of the justification of two judicial decisions and on a problematic recourse to the notion of “human dignity” that are too advantageous for paternalism.

2. Autonomy and Well-being, Selves and Typologies of Interests

2.1 First Step: The Correct Balance of Autonomy and Well-being

If we return to the initial definition of “paternalism”, i.e. if a paternalistic intervention consists in interfering with the freedom of action of a given agent (the agent’s right to decide as she pleases) for her benefit,¹⁴ then we can reconstruct the first analytical step of scenarios of paternalistic intervention as the competition and, often, conflict between two values: autonomy and well-being.

We must therefore distinguish between two main possibilities: 1. that there is, all things considered, a concordance between well-being and autonomy, and 2. that there is, all things considered, a conflict between well-being and autonomy; 3. a third, marginal possibility is that, all things considered, there is a conflict between autonomy and autonomy.¹⁵

¹⁴ These notes embody the primitive notions of many of the most relevant definitions of “paternalism”, such as those by Dworkin (1972, 175), Buchanan (1978, 371), Diciotti (2005, 99), Alemany (2006, 345, 352), and Maniaci (2025, 5).

¹⁵ There is also a fourth possibility that, all things considered, there is a conflict between two components of well-being. I overlook this option because it partially overlaps with the second and third possibilities and, more importantly, can obscure the conflictual nature of the situation. The crucial point to keep in mind is that the two values differ in content, not that

1. If there is all things considered concordance between well-being and autonomy, there is no problem of justification; it will only be necessary to articulate how well-being and autonomy are reconciled, especially when *prima facie* seemed otherwise.
2. On the other hand, if there is all things considered conflict between well-being and autonomy, we must determine, from a consequentialist perspective, which solution should be adopted.
 - 2.1. If considerations relating to the well-being of the individual outweigh or otherwise counterbalance the value of autonomy, the state should *pro tanto* interfere with the individual's freedom of action.¹⁶ To put it another way, in such cases the harm the individual would suffer from her free action is too serious (her action is not proportional), and therefore it is better to interfere with that action.

This appears to be a good reconstruction, for example, of the obligation to wear a helmet when riding a motorcycle in countries where such a mandatory legal rule applies. It makes sense to argue that this obligation serves (i.e., has the function of and is suitable for) protecting the motorcyclist and that such protection is more important than the freedom the motorcyclist might otherwise enjoy, such as the value of feeling the wind in her face, not being overheated by the helmet, or, more generally, the value of deciding as she sees fit. This case, as formulated, is an example of justified paternalistic intervention, and is certainly compatible with liberalism. This is especially true as the interference with the agent's freedom can be deemed minimal and is instrumentally necessary (there appear to be no equally suitable measures less invasive of individual freedom) for her well-being.

In such cases, the motorcyclist, as a conscious, extreme anti-paternalist agent, may ultimately consider that a proper balance is achieved when the

they are similar in kind, as various specifications of the same general value (as, for instance, a utilitarian or another defender of value monism would claim).

¹⁶ I assume the activity of resolving a conflict between values by means of an axiological principle of weighing, in the sense of Chiassoni (2019, 168). As the procedural criteria for this, I use the traditional four proportionality test categories, albeit somewhat loosely.

pleasure of feeling the wind in her face (or similar considerations) outweighs the risk to life and physical integrity. I think there are various reasons to consider such balancing irrational. It is irrational if we focus on the likelihood of the occurring harmful event.¹⁷ It is also irrational if we focus on the balancing of values or the weight assigned to the values themselves.¹⁸

In fact, Western legal systems enshrine various measures with a paternalistic justification, and it can be argued that this justification often does not contradict classical liberalism, as the necessary balance must be achieved and implies only a proportional and sometimes minimal sacrifice of freedom of action, required to improve or, more often, preserve the well-being of the agent affected.

The earlier statement that “considerations relating to well-being outweigh” or “carry more importance” is a somewhat hasty formulation and may appear overly favourable to consequentialism. It can also be argued, in a more deontological vein, that it is a matter of guaranteeing a set of goods considered

¹⁷ For example, suppose the motorcyclist is not extremely prudent and does not drive at 10 km/h on a completely empty road, but is only moderately prudent, driving at 60 km/h on a road where other vehicles sometimes circulate. In such conditions, if she considers the statistical probability of a road accident to be extremely low, we can consider her irrational as she clearly underestimates the likelihood of such an event. Depending on the formulation, this is likely to be a case of statistical fallacy, or an inductive fallacy, concerning either the sample on which the generalisation is based or the generalisation itself. If she disregards the statistical frequency, out of excessive optimism or nonchalance, she could be accused of wishful thinking. Finally, if she does not even consider (at least intuitively, preliminarily) the statistical frequency of the event occurring, she can be deemed epistemically irresponsible. These are some of the possible charges against the motorcyclist concerning statistical frequency. For a different solution to a similar scenario, cf. Maniaci (2025, 52-54).

¹⁸ Indeed, the reproach may emerge from the assignment of a different weight to the two values to be balanced. Consider that the motorcyclist prioritises the experience of freedom (the wind in the face, self-determination) over her well-being (her physical integrity, her life). This can be assessed in an objective or subjective manner. Objectively, by ascertaining that a rational plan of life cannot include a choice involving a significant likelihood of being severely injured, traumatised, or (absolutely irreversibly) killed. Subjectively, by determining that such a possibility is not included in the life plans of most individuals (the percentage of people who, after a period of deliberation and exposure to a minimum amount of relevant information, would choose this is likely to be very low). This argumentation can be based on the agent’s motivational set.

basic, which, regardless of their consequences or competing values, must be incorporated into *any* life plan an individual wishes to pursue.¹⁹

Some authors consider that autonomy (even in the form of freedom of action, which enables autonomy) is not merely another value to be weighed against well-being, but a side-constraint. This could be especially the case if in attributing a kind of lexical priority we are not thinking about any freedom, but the most important ones, such as those pertaining to the Rawlsian first principle of justice (Rawls, 1999, 52 ff.). Despite the apparent soundness of this solution, and the fact that a comprehensive examination of this issue exceeds the possibilities of this paper,²⁰ some perplexities emerged when a strictly deontological perspective on freedom is adopted. In cases such as that of the motorcyclist, we are at least implicitly, and preliminarily, comparing the relative importance of well-being. So, even if we do not put autonomy (or freedom) into a balance with well-being, because we consider that the very balance is precluded, in some sense we are presupposing a sort of comparison between the two values (a “pre-balance”, so to say). Additionally, as I have already pointed out, it is difficult to completely isolate autonomy from well-being, as two discrete values (even if most of my reconstruction of scenarios and thought experiments, for the sake of expository economy, does so). It can be argued that, in relation to the first point, and to a certain extent the second, that only nozickian libertarianism posits the liberty of the will as an absolute value that is not subject to evaluation and, consequently, will prevail over all other values. Even over the well-being of the very owner of that unproblematised freedom. This approach is not sound, as has been demonstrated in part.

2.2. There may also be cases where autonomy outweighs well-being.

For example, in the previous scenario, the motorcyclist does not wear a

¹⁹ Also, if such a framing is plausible, this shows that, alongside the subjective element of satisfying preferences, we can identify an objective element (the so-called “objective list theories”).

²⁰ Specifically, I cannot address here the more fundamental question of whether a conclusive distinction or assimilation between the structures of moral deliberation of a consequentialist sort and those of a deontological sort can be traced. See Bayón (1989).

helmet because she wishes to commit suicide and considers this an ideal way to do so. Nonetheless, the *ceteris paribus* clause must be satisfied: the motorcyclist has made her decision as a result of a reflective process and stable wishes, with all pertinent information being made available to her and without the presence of external coercive pressures.²¹ Again, the *ceteris paribus* clause also translates the assumption that the agent does not incur deliberation flaws, as we will see soon. Should autonomy prevail over well-being, and should the *ceteris paribus* clause be satisfied, it can be concluded that paternalistic intervention would not be justified.

3. In certain cases, autonomy (or freedom) conflicts with another instance of freedom, necessitating its primacy over freedom itself. For instance, the invalidity of voluntary slavery contracts can be conceived as a case of [2.1], or as a situation in which the subject's present freedom is limited to guarantee a future freedom that she could not enjoy if the choice to become a slave were granted (Mill, 1864, 184-185; Maniaci, 2025, 220-232).²²

Therefore, the first condition for a justified paternalistic intervention, or the first step in the theoretical reconstruction of such an intervention, is to proportionally sacrifice autonomy for the sake of well-being. Alternatively, in some cases (such as a voluntary slavery contract), it may involve overlooking a hollow, *prima facie* version of freedom of action or autonomy, in favour of a sound, conclusive version of autonomy.²³

Regarding the consequences of the actions in general, it is sound to establish the principle that, *ceteris paribus*, the more severe is the (risk of) harm, or the more irreversible is harm, the more the rational capacity should

²¹ While the legal obligation still applies, from a critical moral standpoint, it can be argued that the agent's autonomy or other preferences outweigh or otherwise counterbalance those that justify the mandatory use of helmets for the general class of motorcyclists under normal circumstances.

²² This is a case of positive freedom, as I will argue later (§2.5).

²³ It is important to be cautious when extrapolating this reconstruction, which is valid for cases such as voluntary slavery contracts, to other contexts, as it can be used as an apparently liberal approach while serving to disregard autonomy. For a careful discussion of important caveats see Colomer (2004, 176-180).

be required on the agent (apparently or effectively) willing to undergo such a harm (see Feinberg, 1986, 118-121; cf. Alemany, 2006, 404-405; cf. Maniaci, 2025, 43). Concerning the results of the balance, we can conclude that, on a continuum between anti-paternalism and paternalism, the most anti-paternalistic position is the one that gives autonomy lexical priority (and indeed, though this is highly problematic, does so without even considering the countervalue of well-being), while the most paternalistic position is the one that systematically disregards autonomy in order to prevent any harm (even minor) to well-being.

This is undoubtedly only the first step in a complex procedure, and even this initial step must be further qualified. Regarding autonomy, we must determine whether and to what extent consent is valid or could be obtained (that is, what would constitute a correct exercise of autonomy). Regarding both autonomy and well-being, whether and to what extent the conception of well-being is connected to the agent's more stable and most defining preferences, and a more complex or conclusive autonomous "true self". Both issues will be tackled together in the following section.

2.2 Addressing the Problems of Diagnosis and Hypothetical Consent

We should now examine some prominent and recurring ways in which an agent can harm her well-being due to a deficit of rationality. Without excluding analogous situations, and without claiming to be exhaustive, I list situations in which an agent may find herself deciding a course of action and which support the view that the state should *pro tanto* intervene, when the deliberative flaw causes genuine and significant harm (all things considered, if the other conditions are also satisfied).

As we will see shortly, without necessarily adopting a Humean model of individual action – some shortcomings of which will soon become apparent –, it is important to distinguish, in intentional action, between beliefs, desires, and capacity for action. This already provides us with three distinct elements

constituting practical deliberation, each of which paternalistic intervention can affect.

We can establish a taxonomy and a structure of the elements affected by some rationality deficit. Depending on the deliberation component and the severity of the disorder or disorders, and the basic character of the interests involved, a greater or lesser need for state intervention is required.

Concerning the notions of “belief” and “capacity for action”, if taken in isolation and in a rather uninteresting sense (i.e., as lack of information and as lack of factual or normative capacity), these are two assumptions that are not particularly problematic in terms of justification. The state can assist the individual if they lack these two resources: by informing or empowering the citizen, respectively. In this minimal sense, these may not even be considered cases of paternalism: they lack the coercive element and, more fundamentally, they lack the hetero-decision element. In the first case, the state is merely helping the agent to freely make the decision; in the second one, the state is only supporting the decision the agent has freely made.²⁴

More interesting, because evidently problematic, is the case in which desires and beliefs have some defect.²⁵

2.3 Second Step: The Problem of Diagnosis. List of Rationality Deficits

Preliminarily, we must exclude cases of unintentional (not deliberate, conscious) actions. When the issue concerns what the individual desires and believes, we may encounter various scenarios. For example, we can identify the following “disorders”, described in familiar terms in philosophy, which

²⁴ To be sure, even in these cases some salient problems remain, such as how to determine which information is relevant for the agent to take into account (consider, for instance, the epistemic responsibility of the agent herself), or to determine which options should be available.

²⁵ In fact, the Humean model is, at best, limited and its compartmentalising aspirations can be deemed as misleading, if we take into account considerations now commonplace among cognitive psychologists, such as the observation that there do not appear to be any mental states devoid of emotional charge.

often have corresponding manifestations of a clinical (especially, psychopathological) disorder.

Cases of more distinctly emotional problems, or of mixed epistemic and emotional problems, could be the following:

[A] Cases of weakness of will, or *akrasia*. Here, the agent believes that φ is the correct action to take and, to some extent, wants to φ , but is subject to impeding factors that prevent her from φ .²⁶ To this extent, there is no imposition of interests, but rather stimuli towards an already chosen option.

Instances of *akrasia* include severe addiction to substances and behaviours, which can erode the brain's ability to control desires and emotions. For example, narcotics act on the brain's reward system by increasing the quantity and duration of dopamine, causing individuals to refrain from choosing the genuinely preferred alternative option. Another significant case of weakness of will occurs in major depression, due to neural hypoactivity in the prefrontal cortex and hippocampus, which causes individuals not to dare to implement the preferred option.

Of course, minor addictions and depression can be superficial or transitory enough to be negligible as sources of harm to the agent's well-being, while it is cases of major addictions and depression that can be truly paralysing and invalidate genuine choices.

[B] Cases of what has traditionally been called "motivated irrationality", such as self-deception and wishful thinking.

In both cases, the agent – who is normally competent to evaluate evidence and sensible to reasons – holds patently false beliefs, that contradict easily available evidence and does so because of the desirability of the content of those beliefs (see Pedrini, 2013, 4; Mele, 2016, 132).

Self-deception and wishful thinking, in their milder forms, are also widespread phenomena, but among their manifestations we can find truly

²⁶ I am of course simplifying. For a detailed reconstruction, among others, see Kalis, *et alii* (2008).

problematic cases, where the agent's very identity or the main features of her environment are greatly distorted.

Cases of more distinctly epistemic problems could include:

[C] Cases of empirically based inferential problems, such as an erroneous identification of the causal course, the risk involved by the action or one or more of the main effects of one's action.

The aforementioned deficit manifests most acutely in certain individuals diagnosed with psychotic disorders. These individuals often demonstrate an inability to integrate risk-related information, which hinders their capacity to make optimal decisions and which precedes the formation of risk preferences (cf. Purcell, *et alii*, 2022).

[D] Cases of strictly evaluative inferential problems, such as an erroneous balancing of interests.

For instance, not attributing to some interest at the time of action the importance that the agent assigns to it in a more considered deliberation (cf. Alemany, 2006, 137). For more details, I refer the reader above (§2.1), where I critically analyse an example of this kind of error, regarding the extreme anti-paternalist motorcyclist.

It is imperative that some concluding distinctions are made. The first of these relate to the type of condition that could harm significant interests of the holder. The second issue pertains to the nature of the intervention that such a condition might necessitate, namely, an alternative intervention that may be more suitable and less invasive.

It is evident that certain of these deficits (of which a non-exhaustive list is provided here) manifest with such incidence that they do not compromise well-being; in other instances, they occur with a frequency that can be detrimental to the well-being of the agent in question. The diagnosis itself and its impact on decision-making and clinical solutions must be carefully examined. Neither of these deficits needs to be present to the highest degree to be disabling and sustain paternalistic state intervention (for instance, a minor obsessive pattern can paralyse important choices). Conversely, a deficit

at the highest degree cannot necessarily invalidate decisions (for instance, a person with major depression can still adopt appropriate decisions).

Even if the deficit is effectively disabling, this alone does not provide sufficient justification for state intervention in the face of such a rationality deficit. Most notably, it can be argued that paternalistic state intervention is not necessarily the least harmful or most appropriate criterion for addressing and solving these deficiencies in rationality. This might occur *ex post*. In certain cases, alleviating purely circumstantial and transient conditions (e.g., a shock, alcohol intoxication) enables the agent to engage in a cognitive process unencumbered by bias. In other cases, the problem may be alleviated or resolved through psychotherapeutic or pharmacological treatment. Alternatively, patients may be referred for an interview with a doctor who can help them identify and modify their defective psychological processes and certify the soundness of the new unbiased deliberative process (see Maniaci, 2025, 207). Or this might occur *ex ante*. In certain cases, a safety procedure, such as a waiting period before making a decision, can be implemented to ensure the decision is based on an accurate understanding of the relevant facts, the stability of desires and the lack of coercion or compulsion (see Maniaci, 2025, 109, 114-115).

Nevertheless, it is acknowledged that the effectiveness of some of these solutions may be questionable, as certain individuals in certain conditions may be partially or wholly unresponsive to psychological or systemic therapies, or impervious to the control, or safety strategies. Indeed, biases can be pervasive and permanent,²⁷ and self-deception can be so sophisticated that it deceives both the agent and the expert (cf. Maniaci, 2025, 207-208).

²⁷ In fact, as far as I can see, Maniaci's views on the extent of systematic cognitive errors are somehow limited. These errors can sometimes affect even basic cognitive competences, and even experts in those fields (*contra*, Maniaci, 2025, 189-190). Maniaci's rejection of the pervasiveness and severity of biases is based almost exclusively on the rejection of some arguments of Kahneman and Tversky's work (Maniaci, 2025, 187-191), rather than on other studies addressing this issue. To be sure, Kahneman and Tversky's studies in particular can be challenged, and even for reasons beyond those mentioned by Maniaci, such as serious replicability issues (see Schimmack, *et alii*, 2017). Nevertheless, many subsequent studies inspired by these researchers are conclusive, and the wide-ranging and incidence of many

Apart from these considerations, the problematic nature of these cases differs. In fact, we can focus on [B], as it is the most problematic case and, in a certain sense, the paradigmatic one for paternalistic intervention, all for the same reason: such cases can be reconstructed as an internal opposition (not necessarily a contradiction) within the individual. In these cases, the agent appears to be assailed by conflicting beliefs and desires. The idea, as will become evident in the following discussion, is to set the subject *prima facie* against herself and, all things considered, in favour of herself.

2.4 Third Step: The Problem of Hypothetical Consent. The Structure of Well-being and Basic Interests

It is clear from the outset of this work that for liberalism, given the importance it attaches to autonomy, it is crucial to obtain the subject's consent to paternalistic intervention. A fundamental difference between paternalism (justified and unjustified) and perfectionism or moralism is that the latter two positions *ex hypothesi* are not interested in obtaining the consent of the normative recipient. In turn, within paternalism itself a distinction must be drawn between justified and unjustified paternalism, depending on whether consent is (finally) obtained or could be hypothetically obtained or not regarding the paternalistic decision. Preliminarily, if the intervention is based already on the agent's consent, then it lacks the coercive element and, most fundamentally, the hetero-decision element; therefore, for definitional reasons, the decision cannot be considered paternalistic. First, in some cases, e.g. when a transitory affliction recedes, the affected agent mediately gives consent to the paternalistic intervention. Second, sometimes factual consent cannot be obtained, but it is clear that it would be effectively given²⁸ or, more

biases are now widely accepted. Conversely, I am not suggesting that we should place undue emphasis on such errors in studies of paternalism, thereby risking overestimating their significance.

²⁸ Consider, for example, an individual who is aware of all the possible consequences of a debilitating disease is like to be diagnosed with and who leaves a detailed advance treatment statement for the eventuality he contracts the disease. The possibility may arise that the patient changes her decision, but this is quite another problem.

difficultly, that it would be rationally given. In such cases, rational-hypothetical consent is required with the assistance of empirical proof.

If this is the case, we must have some means of identifying the basis and the evidence for the existence and content of that consent; otherwise, we risk undesired effects, such as imposing normative standards on the subject that she does not share, as perfectionism and moralism do, or producing outcomes where any change could be self-confirming, as in cases of brainwashing. Indeed, we should seek a liberal response that does not involve deference to overly demanding justificatory theories (as perfectionism would require: “if the agent were virtuous, she would consent to the intervention”) nor excessive deference to the (partially) irrational subject.

It seems to me that the most suitable strategy for remaining within the bounds of justified paternalism is to refer, in these cases, to the motivational set of the subject involved, i.e. the set of beliefs and preferences the subject actually holds. Two strategies are combined here: the search for the true self and the identification of basic interests.

The first strategy is similar to those of authors who, in cases of individuals affected by persistent and severe addiction to narcotics or severe personality disorders, distinguish two selves: a true self, aiming to develop life plans, and a more superficial self, afflicted by invalidating circumstances (cf. Garnett, 2017, 8 ff.; cf. Husak, 1992, 71 ff.). I cannot pause to examine the merits of these strategies; suffice it to say that they are somewhat metaphorical and can be misleadingly used.

The second strategy involves identifying the subject’s most basic interests.²⁹ The two strategies can be fruitfully combined, by identifying the agent’s true self with her most basic interests. In cases where there is a conflict between basic and non-basic or less basic interests, it is reasonable to give priority to the former. Interests that have special final or instrumental importance in the agent’s life project can be considered “basic”. As I

²⁹ I presuppose Frankfurt’s strategy of stratification of desires and the importance of second-order desires. Cf. Álvarez Medina (2018, 18-19, 44).

anticipated, some basic interests can be found in the Rawlsian primary goods, which are important to any life plan the agent might wish to undertake (Rawls, 1999, 219-220). Other basic interests (which partially overlap in content with the former) are those that shape the agent's self-conception, based on the wishes the agent has deliberately acquired, chosen or modified – or chosen to maintain, even if she could reject them – (cf. Raz, 1986, 290-292) or the idea the agent would like to have of herself, based on second-order wishes – the desires the agent wish to have – (cf. Elster, 1983, 21; cf. Russell, 1999, 71-78). The paternalistic intervention must opt for the true self rather than the superficial one (affected by some invalidating situations), and for the basic interests rather than the less basic ones. Therefore, it can be reasonably posited that, provided we remain within the confines of basic interests, the subject would be in a position to consent or would have compelling reasons (without relinquishing her motivational set) to hypothetically consent to the paternalistic intervention.

How can we obtain actual empirical evidence that the subject would consent, without our argumentative shift becoming an insidious means of imposing interests on the subject that she does not share or will not share? There should be some empirical evidence for hypothetical consent (or, mediately, for the identification of the true self or basic interests): for example, when the subject finds herself tormented by decisions that frustrate what could be considered her basic interests, or when she regrets the options not taken, those that align with the view she wishes to have of herself, and so on. We might, for instance, examine the brain scan or psychological report of a person whose decisions are based on adaptive preferences. To be sure, there may be cases where none of this empirical evidence is present, where the subject fails to recognise (almost) any of her basic interests, such as cases of brainwashing or severe persistent dependence on narcotic substances that radically invalidate executive functions in the subject's life organisation, but these are tragic cases to which the model I am proposing cannot even attempt to respond.

I believe that establishing a connection with the agent's motivational set is fundamental in determining the requirements that must be met for a paternalistic intervention to be justified within the framework of liberalism. This enables the rational deliberative process (embodied in hypothetical consent) to be anchored in some of the individual's preferences, without imposing standards that are entirely foreign to her.

Of course, the content and articulation of the entrenchment of such basic interests must await a new, more detailed formulation. Nonetheless, I wish to make explicit some points of my proposal by distinguishing it from those theories framed in terms of altogether "subjective interests", as opposed to "objective interests".³⁰ As far as I can see, an articulation in terms of "basic interests", some of which are attached to the agent's motivational set, is richer and more theoretically powerful than the alternative strategy, for various reasons.

In general, I do not think the issue is to decide between a completely endogenous set of the agent's preferences and standards (let us call this "subjective", as do theorists such as Maniaci) and a completely exogenous set of preferences and standards, alien to the agent (let us call this "objective", as theorists such as Maniaci do).³¹ Rather, in evaluating a decision as correct, the liberal should seek a normatively relevant element, intersubjectively justified (let us call this "objective"), which must also be anchored in some way to the agent's motivational set (let us call this "subjective"). However, this point remains overly general; I think much attention should be devoted to the ambiguity of the terms "subjective" and "objective" when paired with "interests", to see why in any of these senses (or dimensions), it is reasonable to refer at least partially to the objective element.

³⁰ Maniaci (2025) is a paradigmatic example.

³¹ Maniaci (2025, 4) refers only to "the objective interests" as being completely decoupled from the agent's motivational set and from the agent herself and portrays them as "an ideal interests conception [which], I submit, is prescientific, premodern, deeply deceptive, and absurd" (see also Maniaci, 2025, 5-6). The question is whether the notion of "objective interest" can be used independently of such a radically idealist conception as depicted by Maniaci; of course, my impression is that it can.

Firstly, insofar as we require the agent to follow an inferential path so that their desires and beliefs are rational, not only in a logical-deductive, structural sense (requirement of non-contradiction, and so on) – a “thin theory of rationality”, in Elster’s terminology –, but also in the content of the premises of reasoning themselves (reasoning free from adaptive preferences, what count as a coercive pressure, etc.), we are already entering the realm of the rationally correct, which is material, substantive (in this sense, “objective”), and thus not deferential to the subject (in this sense “subjective”). The latter is clearly what Elster meant when he formulated the “substantive rationality of desires” and identified it with the “right causal formation process”: autonomous desires are only those that have been deliberately chosen, acquired, or modified (see Elster, 1983, 20, 21). So, in the first sense, it is possible to refer, albeit loosely, to the idea of “objective interests” when referring to the normative – whether cognitive or substantive – standards that a community uses to evaluate its members decisions.

Secondly, as I develop at several points (when referring to Rawlsian primary goods and objective list theories), in some sense the very conception of what constitutes the “well-being” that should be assured is, at least partially, objective, in the sense of intersubjectively valid normative standards. Of course, it is not totally objective, if that means excluding any anchorage to the agent’s motivational set.

Thirdly, in some of these cases, we are referring directly to correct decisions and, in fact, often to norms of critical morality. In this dimension and to this extent, it would be inadequate, by definition, to consider that we are not dealing with interests that are in some sense “objective” (at least as part of universally valid norms).³²

The second and third points can be rephrased. Respecting the subject’s decision, which reflects their motivations, is one thing; considering why it is justified (or autonomous, if you prefer), or why it is not justified and why an

³² In addition to what I have pointed out in the note on “moralism”, for these last two points I refer to what I have developed in Fernández Núñez (2023, 414 ff.).

alternative decision and intervention are justified, is another. If we must defer to the decision, we can speak of “subjective elements”, but if the decision can be criticised or, in any case, when it is evaluated, we can refer to “objective elements”. There is a threshold, which we can conveniently call “objective”, as its standards are intersubjectively shared, from which we evaluate the decision to determine whether it is justified.

2.5 A Suitable Interlude. Why Prohibitions and Criminal Sanctions are not the Central Problem. Hetero-decision and Varieties of Coercion

In the introduction, specifically in §1.3, I explained why I regard hetero-decision as the most basic and acute problem that paternalism poses for liberalism. Here I attempt to clarify where the problem does not primarily lie, though it does so at a second step. This concerns the problem of coercion, especially when understood in terms of prohibitions and criminal sanctions. I hope, however, to have indirectly demonstrated a part of this thesis. Indeed, we have not needed to pay much attention to coercion until now when solving the scenarios at hand, but as we approach the background conditions, it becomes evident and more relevant as a problem.

Several insightful contributions about paternalism tend to divert attention from the problem of hetero-decision and focus it elsewhere. In particular, they often emphasise the coercive element, not only in general terms but also in terms of punitive and, moreover, criminal law, suggesting that this is one of the real problems, if not the real one.

This is the case, for instance, with Nino (1989, 420 ff.) and Monti (2019), to cite a pair of contributions that are at least suspicious of paternalism, or Thaler and Sunstein (2009, 4-6), to cite a contribution that instead tries to “sugarcoat” paternalism. The case of the latter authors is particularly significant: paternalism is presented as liberal or libertarian to the extent that nudges, undoubtedly, do not involve coercion. However, the conclusion may change to some extent if we pay attention to what I believe to be the most pressing and basic problem of paternalistic interferences: hetero-decision.

The notion of “hetero-decision” is presupposed by the paternalistic principle I formulated at the beginning of this article: agent *Y* intervenes for the well-being of subject *X* and does so because she has already *decided* for that subject.

The arguments of such authors and my response to them can be formulated as follows: they claim that coercion is the element that distinguishes genuinely or problematic paternalistic interventions from those that are not (which is certainly true),³³ but also that coercion is the real or primitive problem of paternalism (which is false), and particularly – if not exclusively – in the form of sanctions and criminal penalties (which is false). From this, they conclude that the real issue is not so much when and why one agent should decide for another, which is overlooked as the basic problem or at least not considered sufficiently important (which is false).

Instead, the significant justificatory question lies not only, or even primarily, in coercion, and certainly not in punishment: it is quite clear that for liberalism is highly problematic to punish an individual (especially through criminal sanctions) for that individual’s own good.

Of course, coercion and its modalities are highly relevant, for determining either whether something is a paternalistic intervention or whether it is justified. Regarding the assumptions of coercion, a question that easily arises is what conception of “freedom” is adopted when we refer to “interference with freedom” in paternalistic intervention. The tendency or reluctance to identify scenarios as paternalistic or non-paternalistic, or to qualify them as justified or unjustified paternalistic scenarios, can depend, and often depend, on the very conception of “freedom” presupposed.³⁴

Firstly, it is evident that certain authors dealing with paternalism frequently allude to “negative freedom” (in the berlinian sense). However,

³³ It should be recalled that, from the initial definition, coercion is a conceptual component of paternalistic intervention.

³⁴ This is clearly demonstrated by Richter (2021), who compares the definition of “freedom” as presented by liberalism and republicanism, and shows how the republican conception of “freedom” allows much more room for paternalistic interventions than the liberal one.

this cannot be regarded as the sole conception of “freedom” in order to identify paternalistic interventions. Indeed, such a solution would be encumbered by two reconstruction problems. Firstly, if an intervention interferes with freedom understood as “negative freedom”, then, given the broad denotation of negative freedom, there will be many potential interferences. In other words, it is highly probable that the liberal proponent of negative freedom will be more inclined to identify instances of paternalistic interferences much more readily than someone defending an alternative conception (some of which may be more appropriate). Nevertheless, the most problematic issue is not so much the excessive proliferation of situations that can be qualified as “paternalistic”.³⁵ Indeed, the second problem pertains to the defence of negative freedom within liberalism itself. The argument is that maintaining negative freedom as the only possible understanding of “freedom” for liberalism would be too restrictive and, moreover, very unrepresentative, condemning both the problematic approach and the solution to misunderstanding.³⁶ This is because the main problems concerning the *correct* exercise of autonomy within the liberal context affect a freedom understood in the richer sense of “positive freedom”. The ideal situation to consider is that of an agent pursuing her own life plan. Therefore, we are not interested in scenarios of adaptive preferences, where the agent is not genuinely able to choose and is fine to perform actions without interference: consider the traditional example of the slave who is satisfied with her condition and only aspires not to be interfered with in the actions she must perform. The problematic character of such a scenario cannot be grasped with a negative conception of “freedom”, but it can be with a positive one. Instead, we are interested in the scenario in which the agent is master of herself, i.e. she is self-determined, through reflection and her choices, and identifies with what she is (see Celano, 2013, 183-184). Moreover, only by adopting a

³⁵ From a similar perspective, it has been noted that republicanism would tend towards the opposite diagnosis. See Richter (2021); cf. Pettit (1997).

³⁶ This is indeed the main reason why the analysis by Richter (2021) can be considered partial: it is confined to negative freedom as if this were the only sort of liberal freedom.

conception of “freedom” as “positive freedom” can it be understood that certain paternalistic interventions are justified, insofar as they aim to or succeed in promoting autonomy, contrary to what might initially appear. Conversely, certain paternalistic interventions are unjustified insofar as they do not aim to or succeed in doing so.

What I present is particularly clear and crucial in the case of so-called “libertarian paternalism”. It is true that artefacts such as nudges do not exert coercion, but rather influence their addressees.³⁷ However, if they are to be considered paternalistic and problematic as a form of paternalism, it is because, in a sense, they presuppose hetero-decision by the agent who implements the nudges.³⁸ Overlooking this aspect allows the authors to bypass (but only apparently) the problems of the noun “paternalism” and to create the apparently oxymoronic, and in any case paradoxical, expression “libertarian paternalism”.

2.6 *Fourth Step: Some Indications Regarding the Background Conditions*

What I refer to as “background conditions” are the normative conditions favouring intervention, that must outweigh or otherwise counterbalance those against intervention. In this fourth step we are centred in the intervention itself. For instance, intervention is justified if: 1) it is not too invasive; 2) it is not ineffective; 3) it is not too inefficient; and, as a specification of [3], 3’) it is not too costly.

I focus on the invasiveness of the action as a crucial condition. In the previous section we saw that coercion is important. Thus, it is not only a question of whether a measure is coercive, but also of how coercive it is.

³⁷ It can be surmised that in cases such as nudges, an important variable to clarify is what is meant by “influence”, as well as the strength and modalities with which it can be exerted (not to speak of the opacity of the justification or of manipulation). This could be important as a basis for assimilating, at least partially, such “influence” to the phenomenon of coercion.

³⁸ On top of that, the definition of “paternalism” explicitly formulated by Thaler and Sunstein (2009, 5) is itself partial and unduly advantageous to their position, as it includes only the beneficial element, not the coercive element or the most conceptually basic elements of autonomy and hetero-decision.

Consider the state's abstention from interfering with adult's freedom to determine their own diet: it is generally accepted that there are reasons not to interfere because, although such interference may (if sound)³⁹ promote the individuals' well-being, self-determination is considered to outweigh the promotion of well-being.⁴⁰

Some further indications must be added to outline the background conditions more precisely. Adding a nuance to the first condition for justified paternalistic interventions, not only do the ends (like the prevalence of well-being over autonomy, or the other way round) matter, but the means are also crucial in evaluating how severe and invasive state coercion (or eventually influence) can be. Consider the general case of a law containing an exhaustive list of healthy products and requiring that such products be mandatory or highly recommended. In a first, more specific scenario, if the state were to ban all products not included in that list, by prohibiting either their importation or their production, this measure would constitute a very significant interference with individual freedom. In a second scenario, if the state were to impose high direct taxes on those products excluded from the above list, this could be considered a moderately significant interference with individual freedom. Finally, in a third scenario, if the measure took the form of a simple recommendation and resulted in each food shop being encouraged to place the products at eye level we could not speak at all of an interference with individual freedom. Even if the coercive element is absent from the third scenario, unlike the first and second, it could still be deemed a case of paternalism in a loose sense insofar the state is conditioning the norm addressees, and there is a hetero-decision by the legislator (regarding what course of action is better), even if this (meta)decision does not exclude the individuals' final decisions at all.

³⁹ It must be kept in mind that a prerequisite is the assumption that the state knows better than the agent what is good for her.

⁴⁰ The balance may change if the measure is applied to minors: naturally, because the legal system does not take into account, or takes little account of, the self-determination of minors, but also because even if it were argued that it is the self-determination of adults (i.e., parents or guardians) that is considered, the well-being of minors carries greater weight.

In the different scenarios, state action is incisive to varying degrees (and also, among other things, varies in effectiveness), so some interventions can be justified, while others cannot.

3. Anthropology in Law: Competent Agents, Vulnerable Subjects. Critique of Moreso's Proposal

In this second part of the article, I complement, contrast and deepen some initial distinctions (the definitions of the different justifications of a legal norm) and highlight into the importance of the four conditions for justified paternalistic intervention. However, I will do this mostly indirectly, by demonstrating the need to apply such distinctions through the examination of strategies that do not make them and, as far as I can see, would benefit from doing so.

The reconstruction of an anthropology in law is a strategy very similar to the one I have undertaken but much broader in scope. However, it focuses not on the individual recipient of the law, but on the generic subject of the law, understood either as a capable subject or as a vulnerable subject. An even more striking difference, as we shall see below, is the narrowness of my project compared to the breadth of anthropologies in law.

3.1 Pelagius versus Augustine. An Overly Ambitious Project

Depending on what kind of anthropology we defend (that is, depending on how we conceive human rational action), we would allocate more or less space to paternalism: the more capable a subject is, the less conceptual and normative space there will be for justified paternalistic interventions; conversely, the more vulnerable a subject is, the more space there will be for such interventions.

I will examine some recent proposals by José Juan Moreso, which are problematic but very illuminating and stimulating, to develop my argument critically.

Moreso (2021, 131; 2023, 158-159) contrasts two general conceptions of individuals, a contrast he traces back to the theological controversy between Pelagius and Augustine of Hippo.

Pelagius holds an optimistic view of rationality and human action, arguing that agents are responsible for their own actions (in theological terms, “they can obtain salvation by themselves”). In terms suitable for our discussion: actions are fully attributable to agents, who do not need assistance to perform them.

Augustine, on the other hand, argues that Pelagianism is an overbearing conception, as human nature is “vulnerable, wounded, torn, ruined” (*De natura et gratia, caput LIII, 62*). Again, in terms more pertinent to us: subjects are fragile, dependent on others (on God, according to Augustine; on other people according to secular views) and therefore must be grateful to other beings.

John Rawls developed a theory that, according to Moreso, can easily be compared to Augustinian anthropology: there is a clear rejection of the concept of “merit” as the basis for allocating goods in a well-organised political community, and a rejection of the thesis that attributes to personal merit what is actually the result of natural luck and the efforts of others (parents, teachers, and so on). This also implies, as in Augustine, the importance of gratitude: individuals are interdependent, and personal improvement comes from those around them (see Moreso, 2023, 159-160; Moreso, 2025, VII).

As I have pointed out, Moreso’s project is too ambitious and, in a sense, misleading. It is too ambitious because it does not seem reasonable to argue, without further qualification, whether human agents are capable even without the help of others (or whether they are altogether rational or powerful) or whether they are incapable without the help of others (or irrational or vulnerable). It will be necessary, among other conditions, to see in which social context we are examining their actions and identifying the dimensions in which subjects need help, because they are fragile.

Sometimes, however, Moreso presents this anthropology not as a general description or a completed project, but as an idea to be considered within a broader, anticipated project, serving as a counterbalance to the allegedly traditional emphasis on the rationality of human agents (see Moreso, 2023, 169). The first interpretation makes the project overly ambitious and misleading; the second makes it more feasible and sound, to the limited extent we recognise that law and legal philosophers tend to overemphasise human rationality and competence, which is far from self-evident. Therefore, we can consider Moreso's proposal (2025, VIII), viz., the vulnerability element in the notion of "autonomy embodied in vulnerable beings" not so much as a way of "thinking by systems" but rather as an "idea to take into account".⁴¹ However, such an interpretation reduces its explicit ambitions.

This is intended to draw on a realistic anthropology, which is also Moreso's theoretical purposes: an anthropology that is or at least aspires to be partly descriptive (by identifying a deficiency and the causal mechanism linking that deficiency to the possibility of suffering harm) and partly normatively meaningful (by assessing of harm and the desirability of avoiding it, as well as identifying other normative components in the very notions of "fragility", "vulnerability", etc.).⁴²

Therefore, Moreso's objectives and method must also be adjusted to achieve the theoretical and, even more so, practical outcomes that he himself seeks, namely to palliate, to solve situations of fragility and vulnerability.

"Being vulnerable" means, with minimal lexical precision, "being especially susceptible to suffering harm".⁴³ Even without openly

⁴¹ I am employing Vaz Ferreira's (1971, 78-94) distinction. Moreso wrote those contributions as a corrective to the overly Pelagian, to his taste, contributions formulated by Atienza and Maniaci.

⁴² In the last decades, the term "vulnerability" has been used to counter to the assumption of (supposedly excessive) rationality or capacity attributed to the subject. To this extent, I am following to a very widespread usage, although, as we shall see later, it is not without its drawbacks.

⁴³ The literature on the jurisprudential uses of "vulnerability" is extensive. Particularly relevant to my criticism are observations that the concept of "vulnerability" is often used in an opaque manner – as noted by Diciotti (2018, 30 ff.) – or even in a circular way – as noted by Maniaci (2025, 164 ff.).

acknowledging it, as in Moreso's case, we are asserting two theses with significant theoretical and practical implications: [1] The vulnerable subject is unable to decide for herself; [2] the subject's lack of decision or, more often, her erroneous decision is likely to cause her harm.⁴⁴

There is a need to trace both the causes and effects of vulnerability, as well as the problem of diagnosis: to identify the scenarios of vulnerability and, above all, the contexts in which this vulnerability is underpinned by some form of "cognitive fragility". Demonstrating why this is important and how it can occur were the purposes addressed in §2.

If my preceding arguments are sound, the theoretical reasons why these moves by Moreso seem questionable are now quite clear. Moreover, regarding vulnerability, it is important to carefully analyse: 1. the reasons why someone is especially susceptible to suffering harm; 2. the sources of such susceptibility; 3. the causal relation between the susceptibility and the harm; 4. the dimensions and depth in which harm is experienced. For instance, regarding [2] it is interesting to note that Moreso's claims, following some of Rawls' path, seem to be centred on "natural vulnerability", omitting particular or group vulnerability (for the categories see Álvarez Medina, 2021, 75). The practical reasons, on the other hand, include the fact that attributing fragility to an individual can lead to the long-term crystallisation of that fragility and, moreover, erroneously so if the individual was not originally fragile. This results from a series of effects: the individual's self-perception, the perception held by a particular social group, psychological mechanisms that make the situation a self-fulfilling prophecy, and social mechanisms that ascribe an identity to the individual and prescribe a conventional response to that identity (among others, cf. Lukianoff and Haidt, 2018, 19 ff.).

What I have just said pertains to the project of building an anthropology in law but Moreso's conclusion that we need to embrace more cases of

⁴⁴ We can imagine milder forms of these theses: [1'] the agent is unable by herself and without some help (support, supervision), to decide; [2'] through the missed or erroneous decision, the agent does not fully achieve her interests. However, even in such attenuated scenarios, the typical problems associated with the lack of competence persist.

paternalism than classical liberalism commonly accepts is not only a response to his anti-Pelagian anthropology, but also to an argument about the resolution of certain legal cases and the value of human dignity. I will conclude my work by critically examining these two features.

3.2 The False Positives of Paternalism

Moreso's argument is based on a somewhat indeterminate resolution of the legal cases under consideration: both the *Lochner* case and the *Wackenheim* case can be reconstructed not as cases of paternalism, but also as cases of harm to third parties (*Lochner*), perfectionism, and moralism (*Wackenheim*). In the way they are linguistically formulated, I will try to show that the latter is a better reconstruction of these cases.

3.3 Lochner Case: Harm to Third Parties as a Concealed Validation for Paternalism

In *Lochner v. New York*, the US Supreme Court ruled that New York State legislation (the "Bakeshop Act"), which established a maximum of 60 hours per week and 10 hours per day as working hours for bakers, was unconstitutional because, according to the Court, contractual freedom cannot be justifiably limited for the well-being of the contractual parties themselves.

The Court's central reasoning was indeed that such a measure constituted unjustified paternalistic measure. As Moreso quotes:

There is no reasonable ground for interfering with the liberty of person or the right of free contract by determining the hours of labour in the occupation of a baker. There is no contention that bakers as a class are not equal in intelligence and capacity to men in other trades or manual occupations, or that they are not able to assert their rights and care for themselves without the protecting arm of the

State, interfering with their independence of judgement and of action. They are in no sense wards of the State.⁴⁵

Moreso comments:

It is conceivable that a baker might wish to work more than ten hours a day, for example, because he is young and ambitious and wishes to save money to set up on his own. Our political philosophy must show why this measure, which is undoubtedly paternalistic, is justified (Moreso, 2023, 162).

Moreso's treatment of the case is somewhat hasty and, in fact, there are several problems relating to the identification of a paternalistic justification. First, if we reconstruct the *Lochner* ruling with a paternalistic justification, a distinction should be traced between justified and unjustified intentions. As we have analysed, they will be justified if and only if the measure in question: [1] strikes a proper balance between well-being and autonomy; [2] ensures, for instance, a minimum amount of rest for bakers, that is, it compensates for the incorrect balancing or erroneous recognition of values or the way in which they are realised (that is, in general, it remedies some deficit of rationality on the part of the subjects involved); [3] such a resolution is consistent with the basic interests of bakers and with their motivational set. If these conditions are not met, given the coercive nature and benefit to the subject of the intervention, we are faced with an unjustified paternalistic measure.

In any case, aside from this, we may be dealing with a case of harm to third parties. In fact, in the *Lochner* case, it appears particularly clear that we are not so much dealing with a paternalistic justification (i.e., the measure protects each baker from their own decisions) as with a justification based on harm to third parties (i.e., the measure protects bakers who do not want to work at night from those who do).

⁴⁵ *Lochner v. New York*, U.S. 57 (1905).

Thus, there seems to be a problem of strategic interaction that is resolved either by unfair competition from some bakers or by a deterioration in the quality of life for more bakers. Certain bakers who would not work at night on their own initiative feel compelled to do so if other bakers are allowed to work at night. Therefore, it is not so much a question of protecting the latter from themselves,⁴⁶ but rather the former from the latter.

Furthermore, concerning specific mechanism, the idea is that guaranteeing the right to paid holidays, daily rest periods and not having exorbitant working hours as unwaivable would be the ideal mechanism for protecting bakers from employers, to whom they would otherwise be subjected to strong coercive pressures.⁴⁷

Incidentally, the cognitive abilities (more competent, more fragile) that can be assumed of the subjects in each reconstruction of the case are very different.

These two justifications for the case (first, as justified paternalism, and second, as harm to third parties) are absent from both the *Lochner* case itself and the Moreso ruling. However, we must distinguish explicit justifications put forward from implicit justifications presupposed. Indeed, my impression is that the *Lochner* case, as critics of the Supreme Court's decision argue, seems to support paternalism, but especially because harm to third parties, the most notable and solid justification that can be offered for such a case, is only tacitly present and camouflaged as paternalism.

3.4 Wackenheim case: *Perfectionism and Moralism Disguised as Paternalism and the Reappearance of Human Dignity*

While in the *Lochner* case it is easy to find cumulative justifications, two of which show respect for the individual's motivational set, the *Wackenheim* case is more insidious, as the justifications for the court's decision appear

⁴⁶ This misconception is also evident in Moreso (2025, V). This point is made very aptly by Maniaci (2025, XII-XIII; cf. 2025, 20).

⁴⁷ I thank an anonymous reviewer of *Athena* for highlighting this point.

rather alternative and do not show respect for the agents' motivational set. Indeed, if the *Lochner* case was mostly a case of harm to third parties disguised as a paternalistic case to implicitly benefit from the greater legitimacy of harm to third parties for liberals, the *Wackenheim* case is largely a case of moralism and perfectionism disguised as a paternalistic case to implicitly benefit from the greater liberal legitimacy of paternalism.

In the *Wackenheim* case, a person with achondroplasia whose main job consists of being thrown against a wall with appropriate safety measures for the entertainment of the audience, is on trial. The French towns where the show takes place prohibit it, but the court of first instance overturns the administrative act of prohibition. On appeal, however, the French Council of State overturns the first judgement and ratifies the towns council's decision. The case is settled according to this *ratio decidendi*:

The Council of State has ruled that respect for human dignity is a component of public order. Consequently, the authority vested with municipal police powers may prohibit an attraction that infringes upon it, even in the absence of specific local circumstances, by exercising its general police powers (Moreso, 2023, 163).

In this occasion, the textual basis for identifying a substantive justification for the judicial decision is even more laconic than in the previous cases.

One possible justification, which neither the ruling nor Moreso's reconstruction overtly considers, is the harm caused to third parties: that Wackenheim is exposed to public ridicule could undermine not some vague abstract dignity of the group of people with achondroplasia, but the self-conception that many people with achondroplasia have, which would be substantially undermined by practices such as those of Wackenheim, leading to the perpetuation of public denigration of people with achondroplasia as a group.

This resolution based on harm to third parties, very similar to the proposal for *Lochner*, seems to have some merit. Only some merit, however, as it is

necessary to carefully weigh the causal attribution of how Wackenheim's conduct affects other people with achondroplasia and ensure that we do not excessively subordinate the freedom of action of individuals (in this case, Wackenheim) to the will of others.⁴⁸ This is especially important when it comes to an ascriptive characteristic that places the subject in a group, where belonging is not a freely chosen variable.

Incidentally, as in the *Lochner* case, this also presupposes a lesser capacity for self-determination on the part of the normative subject (Wackenheim and anyone who may find themselves in a similar dispute).

The main point is that the Council of State's decision and Moreso's reconstruction offer justification elsewhere, in a much more objectivist direction focused on the notion of "dignity".

Of course, "human dignity" is a very important concept in contemporary legal systems and serves significant functions (among others, Azzoni, 2012). However, there are often several problems with its use.⁴⁹ One of these concerns is the content of such dignity and whether it always prevails over other competing considerations. For example, Moreso, following Atienza, provides a Kantian version of dignity that presupposes a naive compatibilism between dignity and autonomy.

In fact, the issue is that if, as Kant maintains, individual dignity must be respected because it consists in respecting individuals as agents capable of acting according to what they believe to be right (cf. Kant, 2005), are we not assuming that agents do not make mistakes, i.e., that they always choose one of the many options of what can be considered right? The only limit recognised by Kant in this regard is the infringement of the freedom of third parties. As I have tried to show, at least indirectly, this anti-conflictual view of well-being and autonomy is neither realistic (since genuine conflicts arise

⁴⁸ Cf. Maniaci (2025, 171) for additional conditions that a third-party aggression based on "human dignity" must meet.

⁴⁹ For the insidious uses of "human dignity" in paternalistic interventions cf. Maniaci (2025, 167-173).

and, in some cases, the agent is right and in others wrong, without necessarily causing harm to third parties) nor, therefore, theoretically illuminating.

A final problem with human dignity, and the most pressing one for liberalism, is that its use makes it easy to slip into perfectionism and moralism. In fact, in considering a certain conduct “contrary to human dignity”, we often impose on the subject a standard of personal perfection or a standard of moral correctness to which the subject does not adhere. In the *Wackenheim* case, it is genuinely debatable whether Wackenheim himself, whose main job was to be thrown, given the limited employment options available to him, could be persuaded of the fact that his dignity, the ideal to which, if enlightened, he would supposedly subscribe, requires him not to perform that job. As I have already shown, this is a case in which we impose a course of action and certain ideals on the individual against their preferences and, unlike paternalism, we are not even interested in considering their desires and beliefs.

3.5 The Moral of the Story About Alternative Justifications

To sum up, Moreso views *Lochner*’s judicial decision as a case of justified paternalism, whereas I see it as a case of justified application of the harm principle.

Moreso regards *Wackenheim*’s judicial decision as a case of justified paternalism, while I see it as a case of unjustified perfectionism and moralism.

We must think twice about the presuppositions of rationality endorsed by legal norms and judicial decisions. Moreso’s reconstruction of these cases presupposes a lower rational capacity – indeed a generic vulnerability – on the part of the normative parties judged; instead, we can assume higher rational capacities for them. We must await the accurate resolution of the “problem of diagnosis” before providing a solution, and we should be more confident in the rational abilities of the norm addressees. If this is so, we must not be overly deferential to paternalism.

If the detailed determination of the problem of diagnosis is so important for resolving these kinds of cases, *a fortiori*, we should consider attempts to construct an anthropology in law to be an overly ambitious and misdirected project.

References

- Alemany M. (2006). *El paternalismo jurídico* (Iustel).
- Álvarez Medina S. (2018). *La autonomía de las personas* (CEPC).
- Álvarez Medina S. (2021). *La protección de la vida privada y familiar* (Marcial Pons).
- Azzoni G. (2012). Dignità umana e diritto privato, in *Ragion pratica*, n. 38.
- Bayón J. C. (1989). Causalidad, consecuencialismo y deontologismo, in *Doxa*, n. 6.
- Buchanan A. (1978). Medical Paternalism, in *Philosophy & Public Affairs*, vol. 7, n. 4.
- Celano B. (2013). *I diritti nello stato costituzionale* (il Mulino).
- Chiassoni P. (2019). La balanza inesistente, in *Analisi e diritto*, 2019.
- Colomer J. L. (2004). Autonomía y derechos, in J. Betegón, *et al.* (eds.), *Constitución y derechos humanos* (CEPC).
- Diciotti E. (2021). Alcune osservazioni sull'antipaternalismo moderato di Maniaci, in *Diritto e questioni pubbliche*, no. 2.
- Diciotti E. (2018). La vulnerabilità nelle sentenze della Corte europea dei diritti dell'uomo, in *Ars interpretandi*, no. 2.
- Diciotti E. (2005). Preferenze, autonomia e paternalismo, in *Ragion pratica*, 2005/1.
- Diciotti E. (1988). La giustificazione paternalistica di norme, in *Studi senesi*, 1988/1.
- Diciotti E. (1986). Paternalismo, in *Materiali per una storia della cultura giuridica*, n. 16.
- Dworkin G. (1972). Paternalism, in *The Monist*, vol. 56, n. 1.

- Elster J. (1983). *Sour Grapes: Studies in the Subversion of Rationality* (CUP).
- Feinberg J. (1986). *Harm to Self* (OUP).
- Feinberg J. (1984). *Harm to Others* (OUP).
- Fernández Núñez M. (2023). *Derechos e intereses* (CEPC).
- Garnett M. (2017). Agency and Inner Freedom, in *Noûs*, vol. 51, n. 1.
- Hart H. L. A. (1963). *Law, Liberty, and Morality* (OUP).
- Husak D. (1992). *Drugs and Rights* (CUP).
- Kalis A., et al (2008). Weakness of will, akrasia and the neuropsychiatry of decision making, in *Cognitive, Affective, & Behavioral Neuroscience*, vol. 8, n. 4.
- Kant I. (2005) [1785]. *Fundamentación de la metafísica de las costumbres* (Tecnos).
- Lukianoff G. and Haidt J. (2018). *The Coddling of the American Mind* (Penguin).
- Maniaci G. (2025). *The Unbearable Lightness of Legal Antipaternalism* (Springer).
- Mele A. (2016). *El autoengaño desenmascarado* (Cátedra).
- Mill J. S. (1864). *On liberty* (Longman).
- Monti E. (2009). Neutralidad, autonomía y paternalismo, in P. Capdevielle, et al (eds.), *Libres e iguales: estudios sobre autonomía, géneros y religión* (UNAM-III-UBA).
- Moreso J. J. (2025). Foreword: Pelagius or Augustine?, in G. Maniaci, *The Unbearable Lightness of Legal Antipaternalism* (Springer).
- Moreso J. J. (2023). Sobre el semipelagianismo de Atienza, in E. Díaz and F. Laporta (eds.), *Leer a Manuel Atienza* (CEPC).
- Moreso J. J. (2021a). *Lo normativo: variedades y variaciones* (CEPC).
- Moreso J. J. (2021b). Acerca del pelagianismo de Maniaci, in *Diritto e questioni pubbliche*, no. 2.
- Nino C. S. (1989). *Ética y derechos humanos* (Ariel).
- Pedrin P. (2013). *L'autoinganno* (Laterza).
- Pettit P. (1997). *Republicanism* (OUP).

- Purcell J. R., *et al* (2022). A review of risky decision-making in psychosis-spectrum disorders, in *Clinical Psychology Review*, n. 91.
- Rawls J. (1999). *A theory of justice* (HUP).
- Raz J. (1986). *The morality of freedom* (OUP).
- Richter A. (2021). ¿Cuán paternalista es el republicanismo?, in *Analisi e diritto*.
- Russell B. (1999). *Russell on Ethics* (Routledge).
- Schimmack U. *et al* (2017). Reconstruction of a Train Wreck, in *Replicability-Index*, <https://replicationindex.com/2017/02/02/reconstruction-of-a-train-wreck-how-priming-research-went-of-the-rails/>.
- Thaler R. H. and Sunstein C.R. (2009). *Nudge* (Penguin).
- Vaz Ferreira C. (1979). *Lógica viva: moral para intelectuales* (Ayacucho).

ATHENA


CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

Can (Should) Law Do Without Free Will?

ALBERTO ARTOSI

*Former Full Professor of Legal Theory and Logic and Legal Argumentation
University of Bologna (Italy)*

✉ alberto.artosi@unibo.it

 <https://orcid.org/0000-0001-5557-910X>

ABSTRACT

Recent advances in the field of neuroscience are currently casting doubts on the existence of something like free will. This has raised a controversy between those who think that the obliteration of free will is incompatible with holding people responsible for their actions and those who think the other way round. This paper seeks to overcome the problem by investigating the possibility that law could do without free will. An argument to this effect has been advanced by Hans Kelsen in the second edition of his *Pure Theory of Law*. This argument is discussed in sections 2-4 of the paper. The conclusion is that Kelsen's attempt to found law on deterministic grounds does not fit with his legal doctrine, which seems to imply that the answer to the question in the title of this paper is negative. Nevertheless, there is a point in Kelsen's rejection of free will, which can be vindicated. Section 5 introduces and discusses an argument that, although alien to Kelsen's thought, seems to better account for his claim that responsibility does not in any way presuppose free will.

Keywords: Kelsen, free will, legal determinism, imputation, neuroscience

ATHENA

Volume 5.2/2025, pp. 186-203

Articles

ISSN 2724-6299 (Online)

<https://doi.org/10.60923/issn.2724-6299/22488>



*“Imputation presupposes neither the fact or fiction of
[freedom as] causal non-determination”*

H. Kelsen

1. Introduction

Free will has traditionally been considered a presupposition of law necessary for holding people responsible for their acts (should people’s actions be somehow determined, then it would make no sense to hold people responsible for them). Yet, recent advances in the field of neuroscience are currently casting doubts on the existence of something like free will. This has raised a controversy among those who think that the obliteration of free will is not compatible with holding people responsible for their actions and those who think that it is. Both sides are, for opposite reasons, in obvious trouble. The second side, however, can count on a very respectable pedigree. The argument with which, in the 1960s, Hans Kelsen tried to free law from the presupposition of free will can be considered as a paradigm of the “compatibilist” position for both the difficulties it raises, first for his pure theory of law, and the solutions it proposes. In this paper, I will analyse Kelsen’s argument in some detail to answer the question of the title of the paper: whether law can – or even should – do without free will.

2. Kelsen’s Legal Determinism

Kelsen’s rejection of free will as a necessary presupposition of law is found in section 23 of the second edition of the *Pure Theory of Law* (Kelsen, 1967; henceforth PTL):

The assumption... that only man’s freedom (that is, the fact that he is not subject to the law of causality) makes responsibility (and that means: imputation) possible is in open conflict with the facts of social life. The establishment of a normative, behavior-regulating

order, which is the only basis of imputation, presupposes that man's will is causally determinable, therefore not free. For it is the undeniable function of such an order to induce human beings to observe the behavior commanded by the order — to turn norms that command a certain behavior into possible motives determining man's will to behave according to those norms. But this means that the idea of a norm commanding a certain behavior becomes the cause of a norm-conforming behavior. Only because the normative order (as the content of the ideas of men whose conduct the order regulates) inserts itself in the causal process, in the chain of causes and effects, does the order fulfil its social function. And only on the basis of a normative order, that presupposes such causality with respect to the will of the human beings subject to it, is imputation possible (PTL, 94).

The above passage can be said to contain the gist of Kelsen's legal determinism, i.e. the view that to hold people responsible for their actions is possible only based on a normative order which presupposes their causal determination. At this point, to understand Kelsen's view, it is advisable to make some general remarks on the causation of actions.

That actions have causes seems obvious – if they had no causes, it would mean that they happen by chance, in which case we should probably be held even less responsible for them than if they were causally determined. Still, as Nozick (1981) has argued, there is a great difference between having a cause and being causally determined. Suppose that, in a given situation S, person P does action A. If in all situations exactly like S, P had equally done A, A would be causally determined. Suppose instead that P is in a situation in which he can choose to do either action A or action B. Each of these actions has a cause, say C_A and C_B . In the case that P chooses to do A, we can say that A was caused by C_A , but not that it was causally determined by C_A since in the very situation in which he chose to do A, P could have chosen to do B, in which case B would have been caused, but not causally determined, by C_B .

The point is that before P chooses what to do, there is nothing that determines what he will choose; in fact, until the moment in which P chooses to do (say) A, his choice remains open (“free”), i.e. there was the possibility that he would choose to do B.

The distinction between being caused and being causally determined is clear enough. Kelsen, however, was unable to see it because he conceived of causality in terms of the traditional idea of causality as nomic determination, i.e. whenever an event of type A causes an event of type B, there holds a general law (a “law of nature” such as “All metallic bodies expand when heated”) under which any occurrence of A is always followed by an occurrence of B. Applied to actions, nomic determination implies that there is one such law following which “the circumstances before an action determine that it will happen, and rule out any other possibility” (Nagel, 1987, 51). It is hard to think that Kelsen did not realise that embracing such a conception would have jeopardised the very distinction between causality and imputation on which his Pure Theory of Law rests. Indeed, Kelsen’s legal determinism as formulated in section 23 of PTL comes as a surprise to all Kelsenians, including Kelsen himself.

3. Kelsen vs Kelsen

The fact is not only that, as has already been observed (Renzikowski, 2023), there is no comparable text in Kelsen’s previous works,¹ but, what is worse, that we find in the same PTL passages that seem to openly contradict a form of legal determinism such that professed in the above passage. For example, in PTL, 11, we find what can be considered a decisive objection to such a kind of legal determinism:

¹ It would be interesting to investigate the reasons that may have led Kelsen to reject free will and to embrace a strong form of legal determinism. We are probably not far from the truth if we hypothetise the influence of logical positivism and its antimetaphysical stance. On the relationship between Kelsen and logical positivism, see Artosi (2006).

“Validity” of a legal norm presupposes... that it is possible to behave in a way contrary to it: a norm that where to prescribe that something ought to be done of which everyone knows beforehand that it must happen necessarily according to the laws of nature always and everywhere would be as senseless as a norm which were to prescribe that something ought to be done of which one knows beforehand that it is impossible according to the laws of nature.²

The point is so important that Kelsen repeats it again in PTL, 213:

As mentioned in another connection, the possibility of an antagonism between that which is prescribed by a norm as something that ought to be and that which actually happens must exist; a norm, prescribing that something *ought* to be, which, as one knows beforehand *must* happen anyway according to a law of nature, is meaningless – such a norm would not be regarded as valid.

But what is striking here is that the definitive refutation of Kelsen’s legal determinism is found in just three sections after that devoted to the problem of free will:

If the concept of the “ought” is rejected as senseless, then the law-creating acts can merely be perceived as the means of bringing about a certain behavior of men to whom these acts are addressed; therefore as causes of certain effect. One believes, then, to be able to understand the legal order merely as the regularity of a certain course of human behavior... In that case, however, the meaning of an act in which the legal authority commands, authorises, or positively

² This was already asserted by Kelsen with great clarity in the first edition of *Pure Theory of Law* (Kelsen, 1992, 59-60): “Actual behavior corresponds to the system to a certain degree. Complete correspondence, without exceptions, is not necessary. Indeed, there must exist the possibility of a discrepancy between the normative system and what actually takes place within its scope, for without such a possibility a normative system is meaningless; when one can assume that something will necessarily take place, one has no need to order that it happen.”

permits can scientifically be described only as an attempt to create in men certain ideas whose motivating power causes them to behave in a certain way... From this viewpoint “norms” do not exist and the statement that this or that “ought” to be has no meaning, not even a specific positive-legal meaning, different from a moral meaning. From this viewpoint, merely the natural, causally connected events and the legal acts merely in their actuality, but not their specific meaning, are taken into consideration. This specific meaning, the “ought”, is [accordingly] an ideological fallacy and therefore has no place in a scientific description of the law (PTL, 102-3).³

In the light of this passage, any attempt to found the legal order on the causal order of nature is bound to enter into an irresolvable tension with the law’s specific normative meaning. The conflict between Kelsen’s legal determinism and his Pure Theory of Law could not be more manifest. Is it possible that Kelsen did not realise it? Indeed, he was fascinated by determinism to the point of knowingly contradicting what he had claimed in the passage of PTL, 11, quoted above:

Earlier it has been said it would be senseless to issue a norm commanding that something ought to be done of which it is known beforehand that, under a law of nature, it must necessarily always and everywhere take place. This seems to admit that normativity and

³ Substantially the same passage is found in Kelsen, 1992, 32-3: “Normative meaning... is denied altogether from time to time. One views the law, that is, lawmaking acts, solely as a mean of bringing about certain behavior on the part of those human beings to whom the acts are directed; one views lawmaking acts, then, as causes of certain effects. And one believes that in the regularity of a certain pattern of human behavior, one can comprehend the legal system... The assertion (of the legislator or the legal theorist) that whoever steals ought to be punished is regarded, then, as simply an attempt to induce human beings to forbear from theft because others punish thieves; it is regarded as an enterprise to engender in human beings certain notions whose motivating force induce them to behave in an appropriate way... From this point of view, it is causally interconnected natural events alone that are considered, it is legal acts only in their facticity and not the specific meaning that accompanies these acts. This specific meaning – the norm or the “ought” with which the law represents itself and is represented by jurisprudence – turns up as sheer ‘ideology’, and this is true even of meaning that has been refined by the Pure Theory of Law.”

causality are mutually exclusive. However this is not so. The norm that we ought to speak the truth is not senseless, for we have no reason to assume a law of nature according to which men must speak the truth always and everywhere; we know that men sometimes speak the truth and at other times lie. But when man speak the truth or when he lies, then in both case his behavior is causally determined, that means, determined by a law of nature. Not by a law of nature according to which one must always speak the truth or always lie, but by another law of nature, for example by one according to which man chooses that behavior from which he expects the greater advantage. The idea of the norm that one ought to speak the truth can be – in conformity with this law of nature – an effective motive for behavior according to the norm. A norm that would prescribe that man ought not to die would be senseless because we know beforehand that all men must die according to a law of nature. The idea of such a norm cannot be an effective motive for a behavior according to the norm but in contradiction to the law of nature. The idea of such a norm is senseless, precisely because of the lack of the possibility of causal effectiveness (PTL, 94-5).

The least that can be said regarding the above passage (leaving aside any consideration about the particular “law of nature” here involved) is that Kelsen did not take into account the fact that it is one thing to affirm that a norm cannot conflict with a law of nature and a completely different thing to say that its causal effectiveness depends on its conformity with one such law. What is worse, if we connect so closely norms and laws of nature, what would prevent imputation from collapsing on causality? It is quite evident that Kelsen was aware of the problem and that he tried to avoid it by an unexpected conceptual twist.

4. Imputation and Freedom

This conceptual twist is given by Kelsen's recovering of man's freedom at the level of imputation:

[From what has been said so far] It follows that is not freedom, i.e., non-determination of will, but its very opposite, causal determination of will, that makes imputation possible. One does not impute a sanction to an individual's behavior because he is free, but the individual is free because one imputed a sanction to his behavior. Imputation and freedom (in this sense) are indeed essentially linked. But this freedom cannot exclude causality, and does in fact not exclude it. If the assertion that man as a moral or legal personality is free, is to have any meaning, then this moral or legal freedom must be compatible with the causal determination of his behavior. Man is free insofar and because reward, penance, or punishment are imputed as consequence to a certain human behavior; not because this conduct is causally indeterminated, but although it is causally determined. Nay *because* it is causally determined. Man is free because his behavior is an end point of imputation. And this behavior can be an end point of imputation even if it is causally determined (PTL, 98).

To understand this puzzling text, remember that imputation is Kelsen's term for the relation linking a certain behaviour (e.g. stealing) as a condition with a sanction (e.g. imprisonment) as a consequence. In this respect, a legal proposition is analogous to, though essentially different from, a law of nature:

In the rules of law (the sentences by which the science of law describes its object, the law; in contradistinction to legal norms which are prescriptions) a principle is applied which, although analogous to causality, is nevertheless characteristically different from it. It is analogous in that this principle has a function in the rules of law similar to the function of the principle of causality in the

laws of nature... Precisely like a law of nature, the rule of law connects two elements. But the connection expressed in the rule of law has a meaning entirely different from causality, the connection expressed in a law of nature. Quite patently crime and punishment... are not connected as cause and effect. The rule of law does not say, as the law of nature does: when *A* is, “is” *B*; but when *A* is, *B* “ought” to be, even though *B* perhaps actually is not (PTL, 76-7).

Kelsen strongly insists on this last point, e.g. in PTL, 81: “Imputation, implied in the concept of responsibility, is the connection between a certain behavior, namely a delict, with a sanction. Therefore it is possible to say: the sanction is imputed to the delict, but the sanction is not ‘effected by’ (is not ‘caused by’) the delict”; and even more clearly in PTL, 88: “The possibility of the norm being ineffective – that in individual cases it may not be applied or obeyed – must always be present. Precisely in this respect does the difference between legal law and law of nature become apparent.” But if human actions are subject to nomic determination, does it not follow from this that any distinction between causality and imputation, legal law⁴ and law of nature, vanishes?

As we have said, Kelsen’s way out of this difficulty was to recover freedom at the level of imputation: “One does not impute a sanction to an individual’s behavior because he is free,” he writes, “but the individual is free because one imputed a sanction to his behavior.”. But here we realise that Kelsen has ventured onto treacherous ground. For, if, as he maintains, it is the “causal determination of will, that makes imputation possible,” how is it possible that an “individual is free because one imputed a sanction to his behavior”? Maybe we can think that Kelsen was carried away by the desire to translate his idea that causality and freedom are compatible in a rhetorically appealing formula, and that what he meant is instead expressed by the third to last sentence of the passage quoted at the beginning of this section: “Man

⁴ That is “law in the legal sense” (PTL, 79), to emphasize its distinction from the structurally analogous law of nature.

is free insofar and because reward, penance, or punishment are imputed as consequence to a certain human behavior; not because this conduct is causally indetermined, but although it is causally determined.” Indeed, Kelsen’s earlier essay *Causality and Imputation* (Kelsen, 1950) contains the same argument, with almost the same words, but in a more perspicuous form:

It is usual to assume that only [man’s] freedom – and that means his exemption from the principle of causality – makes imputation possible. However, it is just the other way round. Human beings are free because we impute reward, penance, or punishment, as consequence, to human behavior, as condition, not because human behavior is not determined by causal laws but in spite of the undeniable fact that it is determined by causal laws. Man is free because his behavior is the end point of imputation. And it can be the end of imputation even if his behavior is determined by causal law (*ibid.*, 334).

In this form, it is quite clear what Kelsen had in mind: freedom exists – is realized – only at the level of the acts of creation and application of norms, acts which, let us remember, as normative acts are, according to Kelsen, “acts of will,” and therefore free, even if with different degrees of freedom:

To be sure, there is a difference between these two cases [i.e. the creation and the application of norms], but it is only one of quantity, not of quality; the difference is merely that the constraint exercised by the constitution upon the legislator, as far as the content of the statutes is concerned which he is authorized to issue, is not as strong as the constraints exercised by a statute upon the judge who has to apply this statute – that the legislator is much freer in creating law than the judge. But the judge too creates law. And he too is relatively free in this function. For the creation of an individual norm, within the frame of a general norm in the process of applying the law, is a function of the will (PTL, 353).

The very interpretation of the law by the judge is both a cognitive operation and an act of will by which the judge makes a (relatively) free choice among “the possibility shown by cognitive interpretation” (PTL, 354) – a choice that is not available to the addressee of the norm:

If an individual wishes to obey a legal norm that regulates his behavior, that is, if he wishes to fulfill a legal obligation by behaving in a way to whose opposite the legal order attaches a sanction, then this individual, too, must make a choice between different possibilities if his behavior is not unambiguously determined by the norm. *But this is not an authentic choice.* It does not bind the organ which applies this norm and therefore always runs the risk of being regarded as erroneous by this organ, so that the individual’s behavior may be judged to be a delict (PTL, 355; italics added).

The last, definitive, and more metaphysical, reason for assuming human freedom at the level of imputation – “Man is free because his behavior is the end point of imputation” – is rooted in what Kelsen considers “the fundamental difference between imputation and causality” (PTL, 91): in the case of causality the chain of events involved is “endless in both directions” (ibid.); in the case of imputation the chain of events is limited to only two events – the condition and the consequence that is imputed to it in a moral or legal law. Indeed,

Reward, penance, punishment are not imputed to the condition under which a certain behavior is commanded as meritorious or prohibited as sinful or unlawful; they are imputed to the man who behaves in conformity or in conflict with the command, or, more precisely, his behavior in conformity with the command is rewarded, his opposite behavior penanced or punished. In this behavior ends the imputation that constitutes his moral or legal responsibility... The decisive point is: the behavior that, under a normative (i.e., a moral or legal) order, is the end point of an imputation, is, under the causal order, no end

point (neither as cause nor as effect) but only a link in an infinite chain. This, then, is the true meaning of the idea that man, as the subjects of a moral legal order... is “free”. That man, subjected to a moral or legal order, is “free” means: he is the end point of an imputation that is possible only on the basis of this normative order (PTL, 93-4).

There is therefore an end point of imputation – the behaviour conforming or non-conforming with the norm – thanks to which man escapes from the causal process that makes imputation itself possible: this is what the essence of freedom consists of.

In section 1, it was argued that Kelsen’s legal determinism is untenable by his own criteria. This would seem to imply that the answer to the title’s question is negative. Nevertheless, there is a point in Kelsen’s rejection of free will, which can be vindicated. This is the claim that “Imputation presupposes neither the fact or fiction of [freedom as] causal nondetermination.” Kelsen supports this claim with two relatively independent arguments. The first argument is simply that “when imputation is recognised as a connection of facts different from causality but by no means in conflict with it, this fiction becomes superfluous” (PTL, 95). The second argument is that “even a convinced determinist does by no means draw from his view the conclusion that a behavior forbidden by morals or law must not be disapproved or punished – that no imputation must take place” (PTL, 96). In the next section, we will discuss a more refined argument that, although alien to Kelsen’s thought, seems to better account for his claim that imputation does not in any way presuppose the causal non-determination of will.

5. Retributivism

*“A theory of retributive punishment does not await or
depend upon a theory of free will”*

R. Nozick

Nozick (1981) outlined a philosophical theory of retributive punishment which does not presuppose free will. The central assumption is that punishment (re)connects the person who deservedly suffers it with the “correct values” he has “flouted”.⁵

Punishment effects a connection with correct values for those who have flouted them... The system is not one of maximizing the good (connection with correct values) but to eradicating the bad (flouting of correct values) by replacing flouting with linkage. So the role of suffering in punishment is not merely to ensure a significant effect in people’s lives, but... to negate or lessen flouting by making it impossible to remain as pleased with one’s previous anti-linkage (ibid., 384).

The question is: Is this conception of retributive punishment invalidated by the hypothesis that our actions are causally determined? Of course, yes: if our actions are causally determined, then we cannot be held responsible for them and, consequently, we cannot even be punished if they flout correct values by their injustice. As Nozick observes:

It strongly appears that determinism is incompatible with punishment, which raises the philosophical question: given determinism, how is deserved punishment possible. It is not clear that the two are incompatible, but they appear to be, there is a tension

⁵ Indeed, the act of punishment involves “two connections with value... that effected between correct values and (the life of) the punished person, and that between the punisher itself and the correct values when he acts as a vehicle for their having effect” (ibid., 378).

between them. The task is to see how they could fit together (ibid., 394).

Kelsen had posed himself the same question and the same task – to reconcile determinism with imputation – even though, whether he realised it or not, his Pure Theory of Law was not compatible with such a task. Nozick's answer to the question “given determinism, how is deserved punishment possible?” is a sort of “Columbus' egg”: to get rid of the notion of responsibility with all its problematic baggage and to focus on the acts that flout correct values and deserve punishment for those who commit them, whatever their causes:

Retributive punishment effects a link with correct values in those who have flouted them (and who are capable of being linked). By whatever way the person came to be unlinked, came to flout, still, he unlinked and flouting, and so punishment is called for to effect and establish the linkage. “But if it was causally determined that he flout, how can he justifiably be punished for floating?” He is punished for his wrongful act, and he deserves punishment only if it is an act of flouting; we can say shortly... that he is punished for flouting. The punishment establishes the link between him and the values he was anti-linked to; causes of his being anti-linked do not alter the fact of his being so – rather they produce it – nor do they reduce the need for him to be linked (ibid., 395).

What we are faced with here is a clear version of the classical theory of punishment as the restoration of the order of justice⁶. Nozick makes it explicit with the following words:

⁶ Cp., e.g., Hegel, 1963, § 220: “Objectively... by the annulment of the crime, the law is restored and its authority is thereby actualised. Subjectively, it is the reconciliation of the criminal with himself, i.e. with the law known by him as his own and as valid for him and his protection; when this law is executed upon him, he himself finds in this process the satisfaction of justice and nothing save his own act.”

Wrong puts things out of joint in that acts and persons are unlinked with correct values; this is the disharmony introduced by wrongdoing. Punishment does not wipe out the wrong, the past is not changed, but the disconnection with value is repaired (though in a second best way); nonlinkage is eradicated. Also, the penalty wipes out or attenuates the wrongdoer's link with incorrect values, so that he now regrets having followed them or at least is less pleased that he did (ibid., 379).

In turn, Kelsen himself presents his doctrine of imputation as a version of classical retributivism:

retribution is imputation of... punishment to crime. The principle of retribution connects... a behavior which is in conflict with a norm with punishment. Thus it *presupposes* a norm that... prohibits this behavior... just by attaching a punishment to it (PTL, 92).

From this it follows that

According to its inherent meaning [an] order may prescribe sanctions *without regard to the motives* that actually, in each single case, have brought about the behavior conditioning the sanctions. The meaning of the order is expressed in the statement that in the case of a certain behavior – *brought about by whatever motives* – a sanction (in the broader sense, that is, reward or punishment) ought to be executed (PTL, 26; italics added).

The values that, to use Nozick's jargon, are flouted by the behaviour to which a punishment is attached are those established by the norm with which such behaviour is in conflict:

An objectively valid norm according to which a certain behavior 'ought to be', constitutes a positive or negative value. The behavior that conforms to the norm has a positive value, the behavior that does not conform a negative value. The norm that is regarded as

objectively valid, functions as a standard of value applied to actual behavior” (PTL, 17).

It is worth noting that Kelsen’s retributivism is much more radical than Nozick’s. For Nozick, flouting correct values is a question of degree: “Causal determination of action... may lessen the degree of flouting” even if “it does not reduce it to zero; it does not undercut deserved punishment” (Nozick, 1981, 393-4). On the contrary, for Kelsen, flouting positive values does not admit a graduation: “a graduation of an objective value is not possible because a behavior can only conform or not conform with an objectively valid norm, but cannot do so more or less” (PTL, 21). Since the standard of value constituted by an objectively valid norm is a binary standard – conforming/non-conforming – flouting positive values is equivalent to pursuing negative ones and just as entails the punishment associated with such flouting. All that is required for a person to know is what it means to behave in conformity with a norm or to behave oppositely and in this case, what the consequences are. After all, even a serial killer determined to kill by irresistible motives knows that his acts constitute a flouting of the positive values established by the norm that links murder to a justified punishment (i.e., a punishment justified by the norm itself).

6. Conclusions

My aim in this paper was (and is) not to recommend some version of retributivism as a theory of punishment preferable to other theories. Indeed, what I have done is to present two versions of retributivism that reject free will with the intent of showing that rejecting it is just as plausible a philosophical option as accepting it. Of course, this does not mean that the two options are indifferent. To be sure, many (perhaps even most) people will feel uncomfortable with a theory that admits that unjust actions should be punished even if they are causally determined. They are more comfortable

with a theory that assumes that, given free will, the fact that such actions may be causally determined can limit, and in some cases even cancel, their punishability. But the judgement that intervenes here is dictated, more than by a comparative consideration of the intrinsic merits of each option, by our legal and moral sensibility. It is our sensibility that rebels against the idea that we can be punished for our wrongdoings even if we are causally determined to commit – and thus not responsible for – them. But we must not forget that, in the course of a long history, free will has been invoked, or denied, in support of the most diverse causes, starting from its use as an answer to the problem of evil (evil as a consequence of the abuse of our freedom of choice). And throughout this history, it has been the subject of interminable and inconclusive discussions. Whether the progress in the field of neurosciences has contributed (or will contribute) in some way to drying up the “quagmire of questions” (Nozick, 1981, 396) that free will brings with it is, in turn, a question on which I leave it to the experts to pronounce. But even if it were demonstrated that there is nothing like free will, punishing or not punishing actions that flout correct (positive) values would remain a choice that, as Kelsen argued, would bring into play that little or much freedom and responsibility that we are entitled to recognise in ourselves.

References

- Artosi A. (2006). *Kelsen e la cultura scientifica del suo tempo (Kelsen and the scientific culture of his time)* (Gedit).
- Hegel G. W. F. (1963). *Philosophy of Right*, trans. by T. M. Knox (Oxford University Press).
- Kelsen H. (1992). *Introduction to the Problems of Legal Theory. A Translation of the First Edition [1934] of the Reine Rechtslehre or Pure Theory of Law*, trans. by B. Litschewski Paulson and S. L. Paulson (Clarendon Press).

Kelsen H. (1971). Causality and Imputation, in H. Kelsen, *What is Justice? Justice, Law, and Politics in the Mirror of Science* [1950] (University of California Press), 324-49.

Kelsen H. (1967). *Pure Theory of Law*, trans. by M. Knight (University of California Press).

Nagel T. (1987). *What Does It All Mean? A Very Short Introduction to Philosophy* (Oxford University Press).

Nozick R. (1981). *Philosophical Explanations* (Harvard University Press).

Renzikowski J. (2023). No Pure Theory of Law without Free Will, in *Archiv für Rechts-und Sozialphilosophie*, vol. 109, no. 4, 482-96.

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

Peace, War, Democracy: The Value of Different Perspectives

Some Reflections from the VII “Supranational Democracy
Dialogue”

SUSANNA MARIA CAFARO

*Full Professor of European Law, University of Salento (Italy)
Director, Jean Monnet Center of Excellence “EUMANITY DEMOS”*

✉ susanna.cafaro@unisalento.it
ID <https://orcid.org/0000-0002-6431-4207>

EDITORIAL NOTE

Due to the introductory nature of the manuscript, this paper has not been subjected to the double blind peer-review process.

ATHENA

Volume 5.2/2025, pp. 204-209

Conference Papers

ISSN 2724-6299 (Online)

<https://doi.org/10.60923/issn.2724-6299/22432>



What I value more about this hybrid event that was my creature about eight years ago and that, since then, has taken a life of its own, is that I (and we all involved) never stop learning. And the learning process does not necessarily involve understanding something new, but also something we thought we knew, but we see with new eyes.

In 2025, after several months under winds of war, we decided to dedicate the VII edition of SDD to “*EU as a lab in a changing world, citizenship, values and the response to global challenges*”, perfectly conscious that – crisis after crisis – the new challenge was coping with an increasingly insecure world. The topic was, in other words, to explore how our democracies under attack could cope as well as what possible democratic solutions could counter this world’s shift towards the current (un)balance of powers.

The EU offers us – as Europeans - a peculiar perspective, but we are well conscious that other perspectives are needed and welcome in a conversation that, by definition, aims at overcoming borders. One third of this edition’s panellists were, not surprisingly, from outside the EU – as close as from candidate countries or as far as from the other continents (all of them!). A good half of them were academicians, but not law scholars as we organizers, a significant group were representatives of civil society, and a few were international officers.

The success of this edition – with a record of answers to our call for papers – was unfortunately due, in my opinion, to the urgency and the relevance of the topic on the backdrop of the deterioration of global international relations that evokes ghosts of the past.

Those of us who remember (vividly!) the fall of the Berlin wall have witnessed very different historical phases: the phase of bipolarism before that; the one of American unipolarism, and the current scary phase of unstable multipolarism, even if different to the one between the two world wars.

With the German reunification and the collapse of the Soviet Union we in Europe went through a confused transition period, an uneasy one, as the bloody war in Former Yugoslavia testifies, but also a period of hope. The Treaty of Maastricht and the subsequent ones were written with the idea to reunify what was previously East and West in our continent, to overcome the idea of a mere economic integration and move towards “an ever-closer Union”, able to guarantee fundamental rights, a space of justice and even a political ownership. The political significance of the European citizenship (1992), of the EU Charter of Fundamental Rights (2000) and of the EU Treaty title on democratic principles (2007) have not escaped scholars, even if most of public opinion missed it.

On the other side of the ocean, the optimism that reigned after the end of the bipolar world was well exemplified by Fukuyama's 1992 book, *The End of History and the Last Man*, which expressed the concept of the final victory of liberal democracy as the culmination of history and the promise of a peaceful future. He saw global democratization as an inevitable, natural process, the result of economic development and globalization, understood as the universalization of the market economy. In such a world, the conflicts perceived were more cultural than national (see Huntington, 1996), and the main threat to security was terrorism. Yet, there was a clear legal appraisal of what was happening. We still perceived there was much to do; it was not the best of all possible worlds, but it was still (almost) peaceful and seemed to promise further progress.

Unfortunately, since the second decade of the new millennium, the deterioration has been constant. According to all indexes, democracy has dramatically regressed, the polycrisis has fuelled populism and nativism, which in turn have contributed to damage international relations and to hinder the functioning of the European Union just when it was most needed, as a barrier to internal democratic regression and as a bulwark of multilateralism.

In this current phase, we witness the growing geopolitical weight of the BRICS, with an anti-American focus and the antagonism between the United

States and China. With the new Russian imperialism, since the first signs in 2014 to its clear assertion with the invasion of Ukraine in 2022, and with Trump's second term in the United States —accompanied by territorial claims and withdrawal (de facto or de jure) from numerous unilateral negotiations — the world seems to have reverted to the balance of power in which geopolitics prevails over international law. The Israelis' unpunished international crimes in Gaza are just the last drop in a scenario where international order has become an international disorder and where international law is struggling to find supporters, despite the commendable efforts of some.

This scenario, from the European perspective, is well described in the article by Ana Bojinović Fenko and Julija Brsakoska Bazerkoska, through the lens of “actorness”, i.e. the combination of institutional identity, international presence, institutionalisation and capability (a new concept for me as a legal scholar). The Union's responsiveness to the new scenario is analysed in three specific areas: international trade, regulation of digital technology and international conflict resolution. While the new paradigm of strategic autonomy surfaces in all three, this conceptual frame shows very different outcomes in each of them.

From my perspective as a European lawyer, I could imagine such results, but the picture of the scenario and the reasoning by Ana and Julija is different from what mine would have been, just like the picture of the same landscape differs if taken by people in different standpoints. It makes for an interesting reading to provide excellent insights on how the institutional response evolves under duress and how liberal values conflict with the pragmatic, geopolitical approach (and risk to concede).

The article by Esra Akgemci offers another layer of understanding current international relations from a gender perspective. Ample doctrinal evidence is provided on the strict correlation between authoritarian, right-wing populism and anti-gender campaigns so that the rise of both is reflected in foreign policy too. As Esra points out,

authoritarian populists from Thatcher to Trump and Bolsonaro weaponize cultural anxieties, casting LGBTIQ+ people, feminists, and migrants as the enemies of ‘the people’ to legitimise authoritarian rule and a return to a ‘secure patriarchal order’.

This idealized good old patriarchal order is, in a word, *retrotopia*, a narrative as powerful as dangerous, undermining societies as well as international peace.

The answer the author supports, both political and ethical, is in feminist foreign policies. Without countering anti-gender politics and supporting women’s and LGBTIQ+ movements in policymaking, the EU, as well as the democratic countries, deny *de facto* their commitment to democracy and peace.

Lastly, the [post by Luca Belgiorno-Nettis](#) inaugurates a new editorial project aimed at hosting selected non-academic contributions, and beautifully represents the wealth of ideas that civil society can offer in a debate that, to be about democracy, must also be democratic, and therefore open. The founder of the new Democracy Foundation – an Australian not-for-profit research organisation focused on political reform – suggests a different path to peace-making that passes through deliberative democracy. In his well-argued contribution, grounded in academic literature, he offers us the suggestion of a global peace assembly, building on the idea of deliberative peace referendums. The idea that ordinary people sorted by lot may succeed where political leaders fail, because of their attachment to a non-specified national interest or just because of pride, is a fascinating one.

These three perspectives – so different among each of them and so connected in facing the current global challenges through the filter of democratic values – offer a small summary of what is a much broader debate. Accepting being challenged in academic certainties and in thinking outside

of the (many) box(es) we are in is part of the process of listening and understanding. It is the very essence of dialogue.

References

Huntington S. P. (1996). *The Clash of Civilizations and the Remaking of World Order* (Simon and Schuster).

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

European Union's Actorness Amid the Weakening Liberal International Order in the Fields of Trade, Digital Sovereignty and Conflict Resolution

ANA BOJINOVIĆ FENKO

Professor of International Relations, Faculty of Social Sciences, University of Ljubljana (Slovenia)


✉ Ana.Bojinovic@fdv.uni-lj.si

 <https://orcid.org/0000-0003-1896-9269>

JULIJA BRSAKOSKA BAZERKOSKA

Professor of International Relations, Faculty of Law, Ss. Cyril and Methodius University in Skopje (North Macedonia)

✉ j.brsakoskabazerkoska@pf.ukim.edu.mk

 <https://orcid.org/0000-0003-3700-3355>

ABSTRACT

This article analyses the implications for the EU's global actorness stemming from the weakening Liberal International Order (LIO). It elaborates on the autonomous actions that the EU pursues as a response to this particular structural change. The analysis centres on four actorness criteria (Wunderlich 2011), i.e. EU's internal self-understanding and external recognition (both forming the EU's institutional identity), the EU's international presence and the EU's capability (interest, instruments) to establish how the EU adjusted in these elements. To verify this adjustment, we investigate three fields of international cooperation where the EU has developed different levels of its actorness institutionalisation, namely international trade, regulation of digital technology and international conflict resolution. Empirical results reveal that the EU has responded to the weakening LIO by adding a geopolitical dimension to its normative self-understanding, grounding this new identity in a concept of strategic autonomy in all three policy areas. Yet, in the pursuit of an autonomous international action, there is high tension between liberal values and pragmatic competitiveness or geopolitical interest revealed in all three cases. EU's presence and international recognition are higher if this tension is lower, especially when the EU is capable to speak with one voice, such as in the case of trade and resolution of Russian aggression on Ukraine. Contrary, reconciling trade competition, digital innovation and security demands with democracy, multilateralism and respect of international law remains a critical challenge, especially when the EU's capability is internally inconsistent.

Keywords: EU, international liberal order, international trade, digital sovereignty, conflict resolution

ATHENA

Volume 5.2/2025, pp. 210-239

Conference Papers

ISSN 2724-6299 (Online)

<https://doi.org/10.60923/issn.2724-6299/23017>



1. Introduction

The EU started to enact its actorness on the international scene in the form of trade and development of cooperation-based external relations in the 1960s. After the disastrous manifestation of its capability-expectation gap (Hill 1993) in the 1990s, that unfolded in the form of its poor handling of conflict resolution in the post-Yugoslav space, the organisation has increasingly adapted to the major structural transformation of the international system, namely hegemonic rise of the Liberal International Order (LIO) brought by the end of the Cold War. This period represented an opportunity for the EU, which it successfully seized and established a hegemonic position in Europe and in some of the international regimes and areas of global governance. In this way, the EU became a normative power (at least in its neighbourhood), a market power, a civilian power and an environmental hegemon. Although the EU was often estimated to be an “economic giant-political dwarf” (Medrano, 1999), Ginsberg was one of the first to challenge these conventional IR depictions, dismissing the label of “political pygmy” as invalid given the EU’s substantial external political influence (Ginsberg, 2001).

However, since late 2010s America’s fast democratic backsliding, trade protectionism and turn away from multilateralism towards unpredictable bilateralism under both Trump administrations contributed to the intense crumble of the LIO. Another complementary factor, such as Brexit, also weakened EU’s position in the UN Security Council, World Trade Organization (WTO) and in the mainly European centred defence alliance NATO. All these further changed the EU’s hegemonic position amid the weakening LIO.¹ Amid the Russian aggression on Ukraine and Israeli military campaign with elements of genocide in Gaza, the EU’s capability is

¹ LIO is not only challenged by the weakening democratic regime in the USA, but also by the growing autocratization in illiberal democracies (e.g. BRICS group) and also liberal democracies in the very EU (e.g. Chatterjee & Naka, 2022; Wodak, 2019; Nord et al. 2025). This aspect, intensely researched, falls beyond the focus of this article.

no longer self-limited. Even though in the case of sanctions against Russia the EU's willingness has been highly consistent and despite the EU's struggle to speak with one voice on the war in Gaza, it shows a steady effort on denying Israel a position of understanding, let alone support. There is a plethora of studies revealing how the EU as a global actor has responded to international crises (economic and financial, Crimean, migration, Brexit, COVID-19, Russian war in Ukraine and re-manifestation of the Israeli-Palestinian conflict). Conversely, this article tackles the issue of EU adjustment and development of its global actorness amid the weakening LIO, which in the last 15 years became one of the rare permanent features of global order.

In this paper, we focus on autonomous actions that the EU pursues as a response to particular changes in the LIO. As for the selected fields of world politics, we have turned to three areas which have been highly impacted by the crumbling LIO. This process features the abandoning of the value of multilateralism, returning to great power politics or even unilateralism, curtailing the liberal principle of open free trade by "tariff wars," and neglecting the value of international law (Ikenberry, 2018; Mearsheimer, 2019; Hočevár et al., 2023). Lake et al. (2021) demonstrate not only the weakened features of LIO, but also how dysfunctional it has already become. Bojinović Fenko and Brsakoska Bazerkoska (2024) show that, due to changes in world order, the EU actorness adaptations have been to support multipolar international system, multilateral governance, and ground its actions in principles of international (and more specifically, EU) law. Even though the EU has been the pioneer of interregionalism in trade, development and democracy-promotion, it has conducted these aspects of its actorness as complementary to inclusive and universal UN-based multilateralism rather than as an alternative mode of global governance. In its founding and other primary documents, the EU has always supported the UN system (Bouchard and Drieskens, 2013). The EU thus favours a well-functioning multilateral world order, which requires "a certain degree of institutionalisation that

counters unilateral action, limited bilateral solutions, or ill-considered political or military reactions which aggravate sensitive security situations” (Hettne, 2005, p. 560). This directly counters the recent development in the weakening LIO and thus merits research attention.

We analyse the EU’s actorness via basic criteria of actorness as outlined by Wunderlich (2011), namely: institutional identity, international presence, institutionalisation and capability. We focus on how the EU adjusted its global actorness by considering structural changes inflicted on diverse fields of international cooperation, such as international trade, regulation of digital technology and conflict resolution within international peace and security. These are selected as policy areas where the EU has developed three different levels of its actorness institutionalisation. Our aim is to understand how the EU has interpreted and responded to structural conditions in the weakening LIO within these three international fields. First, the paper assesses the issues connected with open international trade, where, under the external dimension of the common commercial policy, the EU has exclusive competences and the longest-practicing actorness. Second, we examine the EU’s actorness on the issues of digital sovereignty, which falls under the external dimension of its common market policy and human rights promotion, where the EU has shared competences with the Member States. Finally, we turn to the international conflict resolution, where the EU has supporting competences within the Common Foreign and Security Policy (CFSP), an area where the EU has faced diverse results in addressing the Russian war in Ukraine since 2022 and the Israeli offensive in Gaza and Lebanon after the Hamas attacks of 2023. Finally, The EU’s normative contributions and actions vis-à-vis contemporary challenges in the three international fields will be examined via legal and content analysis of primary and secondary sources of EU’s origin and from relevant international settings (regimes, international governmental organisations, such as the UN, WTO, the EU primary and secondary legislation and statements by relevant EU institutions).

2. Analysing EU Actorness

Wunderlich (2011) defines actorness by four criteria: institutional identity (i.e. an internal self-understanding or identity and external recognition), international presence (i.e. the capacity to actively influence the external environment), institutionalisation (of which a continuum exists of a lower level – informal institutionalisation and higher level – legalised institutionalisation) and capability (i.e. projection of interests via policy instruments to achieve outcomes). Similarly, these criteria have already been interpreted, although at a more basic level, by Bretherton and Vogler (2006), Ginsberg (1999) and initially by Sjosted (1977). The former defined EU actorness as a capability developed as a response to an opportunity and with a “side effect” of presence – an unintentional influence achieved only by the existence of the EU in the form of third actors’ perception of the EU (Bretherton and Vogler, 2006, pp. 25–27). Most studies of EU actorness have focused on capability, exposing the EU’s gap stemming from internal behavioural limitations in comparison to states (e.g. Sjosted 1977) or originating from external and internal expectations (Hill, 1993).

Bretherton and Vogler (2006) have moved towards structurally-driven analysis of actorness, taking into consideration features of the international system more than internal capabilities. In essence, actorness – theorised as agency or structure driven – necessitates the autonomy of the aforementioned actor. We shall thus operationalise structurally-driven actorness as the EU’s autonomous actions in the following elements of actorness (Wunderlich, 2011): 1. EU institutional identity as an internal self-understanding and as external recognition, measured in EU self-definition in international arenas, and responses to that by relevant international actors. 2. EU international presence, namely its capacity to actively influence the external environment, such as the UN Organisation, the UN system and relevant international regimes. 3. EU capability, i.e. projection of EU interests via policy instruments to achieve outcomes, whereby we will estimate policy coherence

and consistency (EU's ability to speak with one voice and act complementarily in different international policy arenas). We shall take the EU's institutionalisation of a selected policy field as a known element, as EU's legal competences in the selected policy fields (trade, digitalisation and conflict resolution) are known. We present this conceptual model and its operationalisation in Table 1.

FIELD OF WORLD POLITICS		TRADE	DIGITALISATION	CONFLICT RESOLUTION
<i>Element of EU actorness</i>				
<i>EU' institutional identity</i>	<i>internal self- understanding</i>	self-definition and positioning in EU primary law, decisions and strategic documents		
	<i>external recognition</i>	WTO	UN GA debate	UN debate and decisions on war in Ukraine and Gaza
<i>international presence</i>		WTO procedures	trade regime, IOS, ITU, IETF, IEEE	humanitarian and conflict mitigation
<i>capability (interests, instruments)</i>		goals and tools provisioned in EU strategic documents		

Table 1: *Analytical elements and operationalisation of EU actorness*
Source: own upgrade based on Wunderlich (2011)

In the following part, we analyse how the EU has been developing its actorness dealing with three areas which have been highly affected by the crumbling LIO: a) international trade, b) digital sovereignty, and c) international conflict resolution.

3. EU Actorness in the Contemporary Weakening LIO

While immanent individual global crises (i.e. economic and financial, migration, COVID-19) have made the EU initially struggle with coordination and internal unity, these international structural constraints led to greater cooperation within the EU, along with enhanced global engagement. This set the stage for an expanded role in future crises and a rethinking of its global actorness, particularly in health diplomacy and economic resilience. In the process of weakening LIO, the EU's role on the international stage in terms of international trade, regulation of digitalisation and international conflict

resolution was directly challenged by such factors as the continued non-operation of dispute settlement mechanism in the WTO, fast developments in information-communication technology supporting commercialisation of AI, and international conflicts, the Russian war in Ukraine, the Israeli offensive in Gaza and Lebanon after the Hamas attacks, among others.

Moreover, a general development that strongly affected the need for a significant repositioning of the EU's status within the UN and the WTO was the implementation of Brexit in 2020, since the UK holds a permanent seat in the UN Security Council and has a vast global diplomatic and trade presence. While the two sides are no longer part of a common political project, they are deeply interconnected. Therefore, a functional, cooperative relationship is essential for stability, prosperity, and shared influence on the international stage. This notion has become especially relevant with the changing role of the US in the world that became evident with the first and especially the second Trump administration.

3.1 International Trade

After the end of the Cold War, the growing economic and market power of the EU was manifested mainly throughout its WTO membership. When in 1994, the EU became a WTO member, it was perceived as a necessary step for recognition by the international community of the reliability of its own trade policy and manifestation of its presence. The latter was positively enhanced as a result of the EU's decisive role in the development of global trade governance through the creation and development of the WTO and an extensive network of free trade agreements. Today, both the EU and the WTO face numerous challenges and need to reposition themselves with regards to new structural conditions of global trade. It is particularly true in cases of the US pressure with imposing tariffs, the rapid technological changes and the growth of the digital economy, climate change, and the universal values on human rights. This trend in international trade is redefining the EU institutional identity, but also its international presence.

The EU's international presence in the trade arena has been challenged by the strained relationship with the traditionally close ally – the US. Both the first and now the second Trump administration challenged the EU-US partnership and the global trade overall. Challenges to the international trading system had been mounting prior to Trump's first term in office, but the latter was a defining moment, as Trump initiated a more systematic dismantling of norms and rules, while inserting hard geopolitics into the economic realm (Eliasson and Garcia-Duran, 2025). The election of Donald Trump immediately brought a withdrawal of the US from the Transpacific Partnership, a refusal of reappointments or renewals of any WTO Appellate Body member, and a preference for a return to a power-based GATT system (Roberts et al., 2019). The EU relied strongly on the WTO and especially the WTO Dispute Settlement Mechanism to project its views on international trade issues. At present, the EU has 14 ongoing disputes with China and 35 ongoing disputes with the US, while it is represented as a third party in 221 cases – which gives the European Commission a stage to articulate EU perspectives on issues pertaining to international trade.² However, the refused reappointments or renewals of any WTO Appellate Body member by the Trump administration lowers the EU's capacity to actively influence the external trade environment throughout the WTO. The US is redefining the international trade system by insisting on bilateral agreements with the countries around the world, rather than to use the multilateral trading rules imposed by the WTO. Moreover, Trump uses trade tariffs to impose USA interests on the rest of the world. His second term started with the imposition of tariffs on Canada, Mexico and China. In April 2025, his administration imposed a 10% baseline tariff on almost all countries and additional, individualised reciprocal tariffs on countries with which the US run a persistent trade deficit. In mid-July, he threatened to impose a 30% tariff on imports from inter alia the EU, starting on the first of August. According to

² https://www.wto.org/english/tratop_e/dispu_e/dispu_by_country_e.htm.

Trump, the economic security is national security and the goal of the US trade policy should serve to the US to benefit. The Trump decision to revert to “aggressive unilateralism” and pursue a protectionist trade policy, has challenged the EU and completely paralysed the WTO. To sum up, this trend of dismissal of accepted norms, a declining rule of adherence and trade wars, along with geopolitical concerns, have influenced the EU’s international presence twofold: a negative perception of the EU stems from failing as multilateralism, yet a positive perception of the EU stems from its willingness to further value WTO and act itself as a legitimate leader of global liberal trade. In the changed context of international trade within the weakened LIO, especially after the second Trump administration, the EU may be forced to redefine the internal self-understanding.

The EU capability-related reaction to challenges of multilaterally regulated global trade can be observed in the European Commission strategy on trade (2021), which is connected with the EU’s capacity to project its own interests via different policy instruments. It shows how the EU is updating its trade policy. Among other things, there is a strong focus on the review of strategic autonomy, as well as on sustainability and assertiveness (Fahey and Mancini, 2022). The EU’s determination to defend its interests and to ensure that the commitments in its trade agreements are upheld by its trade partners, can be found in the newly institutionalised regime of a Trade Enforcement Officer created in 2020. Furthermore, at the beginning of the Biden administration, the EU together with the US set up the Transatlantic Trade and Technology Council. . The Council set rather high goals to deal with the global challenges in trade and technology with its most significant third country cooperating partner - the US. This new Council represented a new modus operandi for the EU to engage with complex partners, comprising executive to executive engagement, meeting agency counterparts regularly in close groups in an era of EU trade policy deepening its stakeholder and civil society domain overall, leading the EU, like US trade law, to the use of executive-led soft law (Fahey, 2024). Despite the fact that this Council was a

projection of the EU's ability to act coherently in protecting its international trade interests, the Council's overall future is uncertain due to the political changes in the US.

In this context, trade has become a much higher profile policy area for the EU. The EU, as a significant global trade actor, again faces the need to design appropriate responses to international trade conflicts and tensions and to increase its capability. This task seems to be quite complicated given that trade policies increasingly include dimensions of security. In the summer of 2025, the US and the EU agreed on a deal on tariffs and trade, recalling that the transatlantic partnership is a key route of global trade. The agreement came fast and as a reaction to the EU's need to balance US interests in the Ukraine war, which was considered as a decisive moment to conclude the trade agreement. According to the European Council President António Costa, the war in Ukraine was a factor in the EU's accepting its much-criticised trade deal with the US.³ Although the EU is strongly opposed to the return of tariffs, it has experienced escalating tensions with a key ally over tariffs, while the Eastern border is under threat, to be an irresponsible risk. The European Commission President Ursula von der Leyen defended the agreement, arguing it was a conscious decision that avoided a trade war.

The EU's answer to the continuous problem with the US under the Trump administration, which has proven to be a highly unreliable, unpredictable and unstable trading partner so far,⁴ is to adjust itself to the rapid changes of the international trade order. The European leaders unanimously accepted the 15 percent tariff because it is better than the threatened 30 percent, but one certain outcome of this deal that will follow is that it will further incentivise the EU to turn toward a version of economic security, which is cautious of

³ 'Costa breaks ranks on EU-US trade deal, fires warning shot at Trump', *Politico*, available at: <https://www.politico.eu/article/antonio-costa-eu-us-trade-deal-warning-shots-donald-trump/>.

⁴ While the deal was announced on 27 July, and a joint statement with the EU issued on 21 August, less than a week thereafter Trump announced he will impose additional tariffs on all countries upholding digital taxes, legislation, rules or regulations.

the US. The EU will put even more energy into opening other markets and ensuring that they are not dependent on the partner across the Atlantic. Although, in the pursuit of such autonomous international action, there is a high tension between liberal values and pragmatic competitiveness, the EU's presence and positive international recognition are still high, especially since the EU is able to speak with one voice in this particular area. Moreover, in this new era of deeper trade, the role of the EU, as a longstanding proponent of multilateralism, is to offer leadership once again in strengthening and defence of the rules-based multilateral trading system.

3.2 Digital Sovereignty

The area of digital sovereignty development is widely perceived as geopolitically and economically crucial for the coming decades. Technological advances are merging the physical, digital and biological worlds in ways that create both huge promise and a potential threat. The EU has successfully grasped the regulation of the digital aspect of the market and assurance of human rights. As Bradford pithily observes, the EU has been adopting its own human-centric and rights-driven approach to digital regulation vis-à-vis market-centred US strategy and state-centred China's digital sovereignty strategy (Bradford, 2023). The EU's approach focusses on enhancing the individual and collective rights of European citizens and views governments as the entity playing a central role in both steering the digital economy and using regulatory intervention to uphold the fundamental rights of individuals, preserve the democratic structures of society, and ensure a fair distribution of benefits in the digital economy. The EU identifies democracy, fairness, and fundamental rights as key values guiding its policymaking, and are directly engrained in the EU's regulatory instruments with the goal of ushering in a human-centric, democracy-enhancing, rights-preserving, and redistributive digital economy where technology is harnessed for human empowerment (Bradford, 2023). This points to an initial observation that the EU remains highly devoted to its normative self-understanding and has

already gained external recognition of this type of digital sovereignty actor. In this policy field, the EU thus strengthens its institutional identity as a global actor.

As per capability-building, the EU endorsed the EU's General Data Protection Regulation (GDPR) in 2018, and decided on the EU 5G Toolbox to assure cybersecurity governance (Renda, 2021). The EU is the first political community to have produced regulation of AI - the EU AI Act was endorsed in April 2024.⁵ In this document, the above-mentioned self-identification as a global actor stems from EU self-definition vis-à-vis AI as a Union that values and promotes the uptake of human-centric and trustworthy AI, and ensures a high level of protection of health, safety and fundamental rights (AI Act, Preamble, 1). The standard of fundamental rights observation is not defined through the Universal Declaration of Human Rights (the UN GA resolution with a status of general international customary law) but as "enshrined in the Charter of Fundamental Rights of the European Union including democracy, the rule of law and environmental protection, to protect against the harmful effects of AI systems in the Union, and to support innovation" (ibid.). This shows that the EU is aware of its unique understanding of fundamental rights and has linked the use of AI to its fundamental rights-complementary understanding. This gives the EU a strong capability to pursue leadership on the question of global governance in the digital market and AI issues and boosts its positive presence as an element of actorness, yet only with like-minded – liberal democracy-supporting countries. Due to the political importance of AI, the EU response to these emerging technologies reflects broader external policies.

The EU is not an AI industry leader, barring exceptions in given sectors such as industrial robot production. Research, production and marketing of AI

⁵ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.

applications are global endeavours conducted by multinational companies that control the necessary digital and physical assets: data, computing capacity and algorithms. In other words, the EU and its Member States have little direct power to shape AI future trajectory and how it in turn transforms wider society and impacts national and regional interests (Briganti Dini, 2025). The EU's only practicable option at present is to leverage its normative power and internal market, or the Brussels Effect (Bradford, 2020), despite the attacks from the US Trump administration on its tech regulations, arguing they amount to regulatory overreach and censorship. Currently, the EU's success in this field can be measured in external governance of its immanent neighbours, EU membership aspiring states in the Western Balkans. These countries have been under the influence of the Chinese, the US' and EU digital empires. Some of them have joined US pressure to ban Chinese technologies (e.g. Albania, North Macedonia, Kosovo), while others (e.g. Serbia, Montenegro, and Bosnia and Herzegovina) maintain cooperation with Chinese digital companies, often for pragmatic reasons (Vangeli et al., 2025). It is therefore alarming to see that in the countries where EU's LIO supporting stance should have the strongest resonance, the national governments decide to rather hedge between the technological giants at the expense of following fundamental rights protection. Conditionality of liberal norms for extending access to technological access and regulation seems to be a limitation to the EU's positive international presence in digital sovereignty.

Capability-wise, the EU is the weakest on hard technology and digital service provisions. It has relied on American companies for its tech infrastructure, with an estimation that "80 percent of European spending on business cloud and services went to US companies" (Cheslow and Pollet, 2025). In a political debate about whether the EU is capable and willing to break from its dependence on giant US technological companies, in May 2025 it was estimated that creating a European tech stack could cost more than €5 trillion, which exceeds the GDP of the EU's largest economy, Germany, and far exceeds the EU's annual budget (ibid.). A more scientific analysis reveals

that the EU is not doing badly comparatively, but has a very divergent capacity over different fields of the global digital technology stack (Sheikh, 2022). Although the EU already has relevant policies in place for all the layers of the stack, its capacity is very diverse across the layers: it is in dire condition in cloud and application layers, under threat in resource, chips and intelligence layers, it is strong in the network layer, and it displays opportunity in the connected device layer (ibid., pp. 18–19). Despite regulation being a powerful tool at the EU's disposal in global actorness, the EU has not until now used this tool to explicitly strengthen the position of European businesses (Bradford, 2023). Considering this, Sheikh (2022, p. 20) concludes that as such, regulation is most relevant for the EU to help make digital technology stack more in line with European values and to ban certain foreign technologies that are in conflict with those values. He adds (ibid., p. 21) that standardisation is a similarly promising tool that the EU has up to now been successfully employing within different international fora (e.g. International Organisation for Standardisation – IOS, International Telecommunication Union – ITU, Internet Engineering Task Force – IETF and the Institute of Electrical and Electronic Engineers – IEEE).

Together with trade, data governance is one of the key areas of EU action with strong internal and external components. The EU wants to proclaim itself a “global digital player” by being at the forefront of global standard-setting for emerging technologies, while remaining open to trade and investment. Data and privacy regulation are a booming field of law challenging the internationalist standing of the EU, which finds itself having to choose between openness and global projection of its standards (Fahey, 2024). The multilateral efforts by the UN have failed to gain traction, and UN General Secretary Guterres recently concluded that “in the face of the serious, even existential threats posed by runaway climate chaos, and the runaway development of AI without guardrails, we seem powerless to act” (WEF, 2024). However, there is not only a problem striking the balance between liberalisation of trade and the fundamental right to personal data protection,

but also with liberalisation as a LIO norm contesting the very linkage between digitalisation and trade. Concerns about European sovereignty and geopolitical rivalry can clash with Europe's commitment to free global competition (Sheikh, 2022). This is illustrated by the EU's discourse on digital sovereignty encompassing different political areas, e.g. security, economy and right domains (Adler-Niessen and Eggerling 2024). Of these, governance in the economic domain is losing salience to control over security as the EU is shifting from its "traditional" market regulatory orientation to an increasing emphasis on assuring public order, for example in controlling the internet (Flonk et al. 2024) and in digital finance (Donnelly et al. 2024). Therefore, for the EU to strengthen its actorness in the area of digital sovereignty and to take the lead, it needs to find a way to reconcile its normative objectives of liberal trade and data privacy on the one side and strategic geopolitical objectives of state security and market competitiveness on the other.

3.3 International Conflict Resolution

Due to the self-defined image of the EU as a proponent of LIO and supporter of the UN Charter, multilateralism and international law (Treaty on EU – Title V, Chapter 1, Art. 21),⁶ the weakening of the LIO presents an ontological threat to the EU (Adler-Niessen and Zarakol, 2020). In this regard, the EU, which has been a traditional civilian power without direct mechanisms for managing the military conflict, has responded effectively by redefining its capacity to engage in international affairs in the context of "strategic autonomy." Mentioned in 2021 in the European Commission Strategic Foresight Report, strategic autonomy is "commonly defined as the EU's ability to make decisions independently while taking into account its own interests and values," which has been initially developed in the fields of

⁶ Consolidated version of the Treaty on European Union — Title V — General provisions on the Union's external action and specific provisions on the common foreign and security policy — Chapter 1 — General provisions on the Union's external action — Article 21 (OJ C 202, 7.6.2016, pp. 28-29).

economic, technology and trade policy (Steinbach, 2023, p. 974). Ever since international peace and security issues have stepped into this equation of volatile international interdependence, the EU has also brought strategic autonomy into CFSP and more concretely into its Common Security and Defence Policy. We explore the EU's actorness in the cases of Russian aggression on Ukraine since 2022 and Israeli-Palestinian conflict that remanifested in October 2023.

Russia's annexation of Crimea in 2014 and war against Ukraine since 2022 exposed Russia's purposeful ignorance of international law as integral parts of LIO and diminished the value of the multilateral setting offered by the UN to provide "a centre for harmonising the actions of nations in the attainment of [UN] common ends" (UN Charter, Art. 1, pt. 4).⁷ The UN – even in the most narrow understanding as the global system of collective security – seems to be sidelined by these bold displays of power in international politics that replace the open channels of peaceful settlement of disputes and institutionalised multilateral relations. After the US administration's overtures to Russia and its suspension of military assistance to Ukraine, the EU strategic autonomy response additionally encompassed defence as a logically "strategical necessity." Yet, "because the employment of coercive tools in a unilateral fashion questions the legal default of multilateralism and openness" (Steinbach, 2023, p. 973), the EU security capability needs to be conducted in coordination with and complementary to its market capabilities and liberal international norms, e.g. there is a need to focus on strong attributes of civilian role of the military and serve only defensive purposes (Bojinović Fenko and Brsakoska-Bazerkoska, 2024).

In terms of EU presence, from the beginning of the Russian aggression, what can be observed is the EU's unified stance. The EU has only supporting competences within the CFSP when it comes to conflict resolution through the usage of a mix of diplomatic, economic, security, and development tools.

⁷ <https://legal.un.org/repertory/art1.shtml>.

However, the EU international presence and its capacity to actively influence this particular conflict has been quite unified and strong in applying economic sanctions against Russia to achieve political goals and upgrading its enlargement policy. First, the UN system is one of the main international settings where the EU has repeatedly condemned Russia's invasion as a blatant violation of international law, particularly the UN Charter. At the UN, the EU has been a staunch defender of Ukraine and an active voice for multilateralism, international law, and accountability. The EU perceives itself as a defender of the European order, therefore, it projects that view throughout the UN system as well. Although limited by its observer status, the EU plays a coordinating and diplomatic leadership role, particularly through its Member States and alliances within the broader international community. In the UN General Assembly, all EU Member States have supported resolutions that condemn the invasion, call for the withdrawal of Russian troops from Ukraine, reject the illegal annexation of Ukrainian territories by Russia and emphasise the need for a comprehensive, just and lasting peace based on Ukraine's sovereignty. Until the end of February 2022, the US was on the same side with the EU and its Member States in the UN system, yet the US changed stance towards the war in Ukraine and its weakened interest in European security raised the EU's coherence and strategic autonomy approach on this issue.

Second, the EU has remained true to the value of an ever-closer Union of European peoples and amended its enlargement policy outreach to Eastern Partnership countries and Bosnia and Herzegovina that wished to align with EU values. In the wake of a full-scale Russian invasion, Ukraine applied for EU membership on 28 February 2022. In a matter of months, it was formally confirmed by the European Council as a candidate country. Before 2022, for the three of the Eastern Partnership countries - Ukraine, Moldova, and Georgia, EU membership was not offered as a possibility.⁸ This changed in

⁸ See further: Van der Loo G. (2016), *The EU-Ukraine Association Agreement and Deep and Comprehensive Free Trade Area: A New Legal Instrument for EU integration without*

2022 when all three countries applied for EU membership, and in December of 2023 the European Council decided to open membership talks with Ukraine and Moldova. Furthermore, in June 2024 the negotiation frameworks were presented. Georgia was granted candidate status in December 2023, on the understanding that it will take the relevant steps set out in the European Commission recommendations to further advance in the process. Anghel and Džankić (2023), however, exposed a high tension between this EU's quick geopolitically inspired response via enlargement policy boost and the essence of the accession process being grounded in achieving political conditions for EU membership.

As for actorness capability, the EU had showed willingness and capacity to use different policy instruments to respond to such an outstanding international challenge as the Russian aggression in Ukraine, by triggering the activation of the Temporary Protection Directive. The EU used this instrument to step in quickly and efficiently with the Ukrainian refugees, something that was not the case during the 2015 refugee crisis. The Russian invasion of Ukraine triggered the largest refugee situation in Europe in decades, where most people fleeing Ukraine fled to the EU. In the EU, they welcomed the refugees under a temporary protection directive, an instrument that allowed the Ukrainians a visa-free entry. By making the temporary protection available, it helped the EU to avoid disruptions and bottlenecks in national asylum systems. This particular case has shown that if there is a political will and unity among the leaders of EU Member States, the EU can handle large-scale refugee situations relatively well. Finally, the EU is using access to the common market as a tool to punish Russia economically by sanctions. As a CFSP tool, the sanctions to deal with Russia's aggression

Membership (Brill Nijhoff); Emerson M. and Cenuşa D. (2018) (eds.), *Deeping EU-Moldovan Relations. What, why and how?* (CEPS); Emerson M. and Kovziridze M. (2018) (eds.), *Deepening EU-Georgia Relations. What, why and how?* (CEPS); Van Elsuwege P. (2021), The Ratification Saga of the EU-Ukraine Association Agreement: Some Lessons for the Practice of Mixed Agreements, in S. Lorenzmeier, R. Petrov and C. Vedder (eds.), *EU External Relations Law. Shared Competences and Shared Values in Agreements Between the EU and Its Eastern Neighbourhood* (Springer).

against Ukraine are one of the most important and effective mechanisms the EU has at its disposal. Furthermore, the EU provides Ukraine with regular and predictable financial support. It thus once again proves that its economic power and enlargement policy represent the EU's uniquely important external action tools (Cardwell and Moret, 2023). Additionally, in 2025, the European Commission presented three initiatives which were newly established defence capacity-building measures: the white paper for European defence – Readiness 2030, the ReArm Europe plan, the simplification of the Defence Readiness Omnibus and a Defence Readiness Roadmap 2030 to measure progress and discuss the next steps.⁹

In the second case of conflict resolution – the Israeli-Palestinian conflict in Gaza, however, the EU's response has been slow, inconsistent and marked with much less EU presence. After the Hamas attacks on 7th October 2023, the EU's (member states') reaction to the Israeli offensive in Gaza and Lebanon varied. The EU has expressed a dual stance in response to the Israeli offensive following the Hamas attacks. While reaffirming Israel's right to self-defence, EU officials have also called for the protection of civilians and strict adherence to international humanitarian law. The statement on behalf of the EU by the High Representative released the very same day of the terrorist attacks condemned "in the strongest possible terms the multiple and indiscriminate attacks across Israel by Hamas," and called "for an immediate cessation of these senseless attacks and violence, which will only further increase tensions on the ground and seriously undermine Palestinian people's aspirations for peace."¹⁰ European leaders have repeatedly urged both parties to exercise restraint, deescalate the situation, and avoid further exacerbating of the humanitarian crisis in Gaza and Lebanon. Furthermore, the EU has

⁹ https://commission.europa.eu/topics/defence/future-european-defence_en.

¹⁰ Council of the EU, "Statement by the High Representative on behalf of the European Union on the attacks against Israel", 7 October 2023. Available at www.consilium.europa.eu/en/press/press-releases/2023/10/07/statement-by-the-high-representative-on-behalf-of-the-european-union-on-the-attacks-against-israel/.

emphasised the need for a swift return to dialogue to prevent the conflict from spiralling into a broader regional crisis.

The Israeli-Palestinian conflict manifestation exposed divisions within the EU which existed long before the 7th October 2023 and based on which the EU's actorness capability on this issue remained limited. The initial solidarity gave way to familiar disagreements and revealed the differences in national interests and perspectives on the conflict. Some of the EU states, like Germany, Austria, and Czechia, emphasised unwavering support for Israel, while others, like Ireland, Spain, Belgium and Slovenia, were more critical of the scale of Israel's military response and advocated stronger protection for Palestinian civilians. These divisions made it difficult for the EU to adopt a fully unified stance on the conflict, humanitarian aid to the Palestinians and the issue of Palestinian statehood. The EU Member States had consensus on only one thing – the need for deescalation. The EU's mitigation strategy, employed via diplomatic, selective trade and humanitarian instruments, did not increase the EU's capability in terms of internal consistency and the EU's international presence was limited to the Quartet and a few other multilateral platforms (Akgül-Açıkmeşe & Özel, 2024).

The divergent vote of EU Member States in the UN confirmed these deep divisions within the EU. On 27th October 2023, the UN General Assembly adopted Resolution ES-10/21 calling for an immediate and sustained humanitarian truce and cessation of hostilities, while condemning all acts of violence aimed at Palestinian and Israeli civilians and demanding all parties to immediately and fully comply with their obligations under international law. Only seven EU Member States voted in favour, while four of the 14 members that opposed the resolutions were EU Member States –Austria, Croatia, Czechia and Hungary, while most EU Member States abstained. The EU position in the UN General Assembly and the following resolutions on this issue, send a message that when it comes to the Middle East, a united European Foreign Policy is an illusion (Soler, 2024). The EU stance in the UN system regarding the cease-fire and the inconsistent messages addressing

the Israeli leadership are just some of the ways in which the EU exposed its internal divergences.

Another element of the weak EU actorness capability visible well before the attacks on the 7th of October 2023 was the question of not whether, but when to recognise Palestine as a state. Whilst EU member states agree on the goal of a two-state solution, they disagree significantly on how it should be achieved. Vignoli et al. (2025) identify three like-minded groups.¹¹ Sweden was the first Western European country to recognise the State of Palestine in October 2014. This recognition was driven by the belief that acknowledging Palestinian statehood would help level the playing field in peace negotiations and encourage a more balanced dialogue between the two parties (Soler, 2024). The scale of the humanitarian crisis provoked by the Israeli militarys retaliation in Gaza following the October 2023 Hamas-led attacks, pushed other European countries to recognise Palestine as a state. The lack of a unified position on Palestinian statehood among Member States presented a serious challenge for the EU to achieve stronger external recognition as a single-voice actor in conflict resolution. Spain and Ireland, together with Norway (not an EU Member State), synchronised their decisions to recognise Palestinian statehood in May 2024, followed by the Slovenian recognition a few weeks later. In the absence of the EU common position on Palestinian statehood, in May 2024 the EU Member States were once again divided in their voting regarding the UN General Assembly Resolution aiming to press the Security Council to give favourable consideration to a full Palestinian membership. The resolution won a majority of 143 votes in favour, including the votes of some EU members that had not yet recognised Palestine, such as Belgium, Greece, Portugal, and quite significantly, France. Many EU Member States abstained the vote, while Hungary and Czechia voted against

¹¹ Sweden, Ireland, Belgium, France and Spain have consistently criticised Israel for its military occupation of Palestinian territories and the settlement expansion. The second group includes Austria, Czechia, Hungary, Poland, Slovakia or the Baltic states, who tend to support Israel. The third group are countries like Germany and the Netherlands, which balance between pro-Israeli and pro-Palestinian views (Vignoli et al., 2025, pp. 3–4).

it. This can be assessed as a positive phenomenon in terms of the EU's overall international recognition, although it does not yet resonate as an EU collective effort, but rather as individual member states' actions.

Developments in 2025 within the EU regarding recognition of Palestine have been intense. A European Parliament resolution in early September called for Palestinian recognition to enable a two-state solution.¹² At the UN General Assembly in September 2025, together with Canada, UK and Australia - France, Belgium, Portugal, Luxembourg, Malta and Andorra declared the recognition of the State of Palestine. These historic decisions brought the number of countries that recognised the Palestinian state up to 156 - 15 out of 27 EU Member States. These developments indicate that the EU is building its capability to speak with one voice also in the case of recognition of Palestinian statehood. Additionally, the EU has shown its diplomatic capacity to align neighbouring non-member countries to its foreign policy position.¹³

At the same time, the bloc is assessing its options after finding Israel in breach of human rights obligations under the Association Agreement. The EU-Israel Association Agreement is a deal that opens up advantageous bilateral relations. Suspending the agreement and all preferential trade with Israel, the EU has a chance to make a difference in the Middle East. The EU has recently started to discuss the sanctions against Israel, as the European Commission proposed on 17 September 2025 suspension of trade concessions and imposition of sanctions on extremist ministers of the Israeli government and violent settlers.¹⁴ There are instruments at the EU's disposal that can strengthen its international presence and geopolitical influence in the Israeli-

¹²<https://www.euronews.com/my-europe/2025/09/11/european-parliament-calls-for-recognition-of-state-of-palestine>.

¹³ Several European states have aligned themselves with EU's statement on the first phase of the Comprehensive Plan to end the Gaza Conflict put forward by President Trump in early October 2025, namely: Albania, Armenia, Bosnia and Herzegovina, Iceland, Georgia, Liechtenstein, Moldova, Montenegro, North Macedonia, Norway, Serbia, and Ukraine (Council of the EU Statement, 2025, 10 October).

¹⁴https://north-africa-middle-east-gulf.ec.europa.eu/news/commission-proposes-suspension-trade-concessions-israel-and-sanctions-extremist-ministers-israeli-2025-09-17_en.

Palestinian conflict in Gaza, but in this case, the EU still lacks the consensus to act more decisively. We can observe, however, that in the case of conflict resolution and related severe cases of human rights violation, the EU's arguments for global actorness are not a dilemma between security or economy-related strategic interest trade-offs, but rather historically-related issues of individual member states.

4. Conclusions

The article elaborated on the international actorness that the EU managed to pursue as an autonomous response to the weakening LIO. We employed conceptualisation of EU actorness and constitutive elements of its institutional identity, that is manifested as internal self-understanding and external recognition, international presence and capability. Findings are summarised in table 2 below.

We established that the EU's internal self-understanding shifted from a normative power rooted in multilateralism to a more assertive and strategically autonomous actor, as evidenced by its updated trade strategy and institutional innovations, such as the Trade Enforcement Officer and the Transatlantic Trade and Technology Council. Yet at the same time, in the field of digital sovereignty, the EU keeps asserting its role as a normative power committed to human-centred and rights-based governance. Through landmark regulatory instruments, such as the GDPR, the 5G Toolbox, and the AI Act, the EU articulated a distinct identity rooted in democracy, fundamental rights, and the rule of law. However, the EU response to Russia's aggression against Ukraine and the Israeli offensive in Gaza reflects a deepening of its internal self-understanding not only as a normative, but *also* a geopolitical actor. In this context, the EU has recently come to understand itself as a defender of European order, especially since it can no longer count on a liberal defensive alliance with the US as a guarantee for regional peace and security.

<i>Element of EU actorness</i>	TRADE	DIGITALIZATION	CONFLICT RESOLUTION
<i>internal self-understanding</i>	clashes between open liberal market and strategic autonomy	clashes between human rights-based provision of digital services and strategic autonomy and competitiveness	clashes between civilian power, UN and international law support and autonomous security and defence provision
<i>external recognition</i>	<i>in the WTO:</i> normatively strong, but limited by US-China rivalry	- <i>in the UN:</i> strong, potential for further “Brussels effect” - <i>in the neighbourhood:</i> unpersuasive	- <i>in the UN:</i> strong on Ukraine, weak on Gaza - <i>in the neighbourhood:</i> pragmatic, geopolitical
<i>international presence</i>	strong in WTO, FAO	standard-setting and regulation in IOS, ITU, IETF, IEEE	fast and strong on Ukraine in the UN and NATO, only like-minded states on Gaza
<i>capability</i>	Strategy on trade, Trade Enforcement Officer	- <i>strong:</i> GDPR, EU 5G Toolbox, AI Act - <i>weak:</i> resource, chips and intelligence - <i>very weak:</i> cloud and applications	- <i>high consistency:</i> on Ukraine (sanctions, refugee intake, UN representation), <i>mid-level consistency:</i> on Gaza <i>strong capability-building:</i> defence

Table 2: *Elements of EU actorness in trade, digitalisation and conflict resolution*

The EU's external recognition as a reliable and principled trade partner has been challenged by the unpredictable US tariff manoeuvres and the influence of China's rapid growth and internationalisation model. A limitation of the EU's recognition, however, remains its inability to reinvigorate multilateral dispute settlement within the WTO. The EU regulatory leadership in the field of AI has earned global attention, continuing the “Brussels Effect,” whereby its standards influence third countries and transnational companies. In its support for Ukraine, the EU has gained recognition as a principled and responsive actor, especially within the UN General Assembly. Its swift humanitarian and financial assistance, sanctions against Russia and activities in the UN have reinforced its image as a defender of liberal international norms. However, EU's recognition also has limitations. In the Western Balkans, for example, the competing digital empires of the US and China challenge the EU normative influence in the countries aspiring to the EU

membership. Additionally, the EU's inconsistent response to the Israeli-Palestinian conflict in Gaza, sidelining by the US and the lack of a unified position on Palestinian statehood reveal a serious challenge for the EU to achieve stronger external recognition as a single-voice actor in conflict resolution. The case of contested normative leadership in its very neighbourhood and inconsistency of internal policy are quite meaningful limitations to recognition of the EU as a credible protector of LIO.

In terms of international presence, the EU remains a central player in global trade governance, actively negotiating agreements and responding to US-initiated protectionism with agility. In digital governance, the EU is increasingly visible, though it displays diverse capabilities in different layers of the global digital technology stack. While the EU lacks industrial leadership in AI development, cloud service and applications, its proactive regulatory stance and institutional innovations have positioned it as a key player in the international regime of digital technology global standard-setting. In conflict resolution, however, the EU's presence is rather unbalanced: it is strong in the case of war in Ukraine and only gaining sight of a possible unified actorness in case of conflict in Gaza. In the UN institutions, the EU has been influential, particularly in condemning Russia and advocating for respect of international law. However, the conflict in Gaza exposes the limits of the EU's international presence due to low capability caused by internal inconsistency: although the number of like-minded states on Palestinian recognition is growing, the EU is not able to bring Israel to responsibility in cases of multiple violations of international law.

Finally, the EU's capability, both in terms of the ability to determine its interests, develop and effectively apply foreign policy instruments, is increasingly defined by its pursuit of strategic autonomy. By diversifying trade partnerships, reinforcing the negative impact of sanctions, and integrating security considerations into trade policy, the EU is positioning itself to respond effectively to the protectionism-leaning US trade policy and to punish Russia for aggression. Through economic sanctions and diplomatic

engagement, the EU has demonstrated its capacity to shape global responses to Russian aggression, whereas trade-related sanctions against Israeli breaches of international humanitarian law are still under consideration. In the digital agenda, the EU must reconcile competing objectives, such as openness in trade, protection of personal data, and digital security, whereby it must balance between liberalisation and control. The EU has shown growing capability and interest in asserting strategic autonomy, especially in security and defence-related issues. Instruments such as the CFSP sanctions, enlargement policy, and financial aid have proven effective and timely in the case of war in Ukraine, yet they have not been agreed upon in the case of Israeli war in Gaza.

Empirical results reveal there are slight differences between the three policy areas, which are grounded in different EU competence. In the field of trade, the European Commission applies its instruments quite effectively, whereas policy such as digitalisation demands internal alignment with areas of EU exclusive competence, for instance, such as market, trade, competition, as well as EU supportive competence. Namely, industry policy, consumer protection and areas of freedom, security and justice. In the end, CFSP-based conflict resolution is the weakest in the EU list of competences, yet this is also the area where the EU has just started its capability-building. The latter includes not only defence investments and strategic planning, but also slow, yet continuous foreign policy alignment on the Israeli-Palestinian conflict. The EU has responded to the weakening LIO by adding a geopolitical aspect to its normative self-understanding, grounding this new identity in a concept of strategic autonomy in all three policy areas. Yet, in the pursuit of such autonomous international action, there is high tension between liberal values and pragmatic competitiveness or geopolitical interest revealed in all three cases. The EU's response to Russia's war in Ukraine stands as the only exception to this rule. When this tension is lower, EU presence and international recognition are higher, especially when the EU is able to speak with one voice. In contrast, reconciling the difficulties of marrying trade

competition, digital innovation and security demands with democracy, multilateralism and respect for international law remain a critical challenge, especially when the EU's actorness capacity is low due to internal inconsistency.

References

- Adler-Nissen R. and Zarakol A. (2021). Struggles for Recognition: The Liberal International Order and the Merger of Its Discontents, in *International Organization*, vol. 75(2), 611–634.
- Adler-Nissen R. and Eggeling K. A. (2024). The Discursive Struggle for Digital Sovereignty: Security, Economy, Rights and the Cloud Project Gaia-X, in *Journal of Common Market Studies*, vol. 62(4), 993–1011.
- Akgül-Açıkmeşe S. and Özel S. (2024). EU Policy towards the Israel-Palestine Conflict: The Limitations of Mitigation Strategies, in *The International Spectator*, vol. 59(1), 59–78.
- Anghel S. and Džankić J. (2023). EU enlargement in a geopolitical era: Between acceleration and ambiguity, in *Journal of European Integration*, vol. 45(6), 789–805.
- Bouchard C. and Drieskens E. (2013). The European Union in UN politics, in K. E. Jørgensen and K. V. Laatikainen (eds.), *Routledge Handbook on the European Union and International Institutions* (Routledge), 115–127.
- Bojinović Fenko A. and Brsakoska-Bazerkoska J. (2024) The EU as a Global Actor: The Significance of Changes in the World Order From 2004 to 2024 as Regards EU Actorness, in *Studia Europejskie – Studies in European Affairs*, no. 2, 7-26.
- Briganti Dini G. (2025). The EU's Response to the Fragmented Emergence of Artificial Intelligence, in O. Costa et al. (eds.), *EU Foreign Policy in a Fragmenting International Order* (Palgrave-McMillan), 207-233.
- Bradford A. (2023). *Digital Empires: The Global Battle to Regulate Technology* (OUP).

- Bradford A. (2020). *The Brussels Effect: How the European Union Rules the World* (OUP).
- Bretherton C. and Vogler J. (2006). *The European Union as a Global Actor*, 2nd (Routledge).
- Cardwell P. J. and Moret E. (2023). The EU, sanctions and regional leadership, in *European Security*, vol. 32, no. 1, 1-21.
- Chatterjee M. and Naka I. (2022). Twenty years of BRICS: Political and economic transformations through the lens of land, in *Oxford Development Studies*, vol. 50(1), 2–13.
- Cheslow D. and Pollet M. (2025). Europe wants to free itself from US tech. Can it? *Politico*, 5 July 2025, <https://www.politico.com/newsletters/digital-future-daily/2025/05/07/daniella-euro-stack-00333636>.
- Council of the EU. (2025, 10 October). Israel/Palestine: statement by the High Representative on behalf of the European Union on the Comprehensive Plan to end the Gaza Conflict. <https://www.consilium.europa.eu/en/press/press-releases/2025/10/10/israelpalestine-stateme-nt-by-the-high-representative-on-behalf-of-the-european-union-on-the-comprehensive-plan-t-o-end-the-gaza-conflict/>.
- Donnelly S., Ríos Camacho E. and Heidebrecht S. (2024). Digital sovereignty as control: the regulation of digital finance in the European Union, in *Journal of European Public Policy*, vol. 31(8), 2226–2249.
- Eliasson J. and Garcia-Duran P. (2025). EU Trade Policy in Light of a Fragmented Liberal International Order, in O. Costa et al. (eds.), *EU Foreign Policy in a Fragmenting International Order* (Palgrave-McMillan), 27-55.
- Fahey E. and Mancini I. (2022). Introduction: understanding the EU as a good global actor: whose metrics?, in E. Fahey and I. Mancini (eds.), *Understanding the EU as a Good Global Actor - Ambitions, Values and Metrics* (Edward Elgar Publishing), 1-17.
- Fahey E. (2024). Legal convergence through soft law? The EU-US trade and technology council (TTC), in *European Foreign Affairs Review*, vol. 29(4), 471–492.

- Flonk D., Jachtenfuchs M. and Obendiek A. (2024). Controlling internet content in the EU: towards digital sovereignty, in *Journal of European Public Policy*, vol. 31(8), 2316–2342.
- Ginsberg R. H. (2001). *The European Union in International Politics: Baptism by Fire* (Rowman and Littlefield).
- Ginsberg R. H. (1999). Conceptualizing the European Union as an international actor: Narrowing the theoretical capability–expectations gap, in *Journal of Common Market Studies*, vol. 37(3), 429–454.
- Hill C. (1993). The Capability-Expectations Gap, or Conceptualizing Europe's International Role, in *Journal of Common Market Studies*, vol. 31 (3), 305–328.
- Hočevár M., Rutar T. and Lovéc M. (2023) (eds.). *The neoliberal world order in crisis, and beyond: an East European perspective* (FDV).
- Ikenberry J. G. (2018). The end of liberal international order?, in *International Affairs*, vol. 94 (1), 7–24.
- Lake D. A., Martin L. L. and Risse T. (2021). Challenges to the liberal order: Reflections on international organization, in *International Organization*, vol. 75(2), 225–257.
- Mearsheimer J. J. (2019). Bound to fail. The rise and fall of the liberal international order, in *International Security*, vol. 43(4), 7–50.
- Nord M., Altman D., Angiolillo F., Fernandes T., Good God A. and Lindberg S. I. (2025). *Democracy report 2025: 25 years of autocratization – Democracy trumped?*, V-Dem Institute, University of Gothenburg. https://v-dem.net/documents/60/V-dem-dr__2025_lowres.pdf.
- Renda A. (2021). Securing 5G: The EU Toolbox and the challenge of coordinated cybersecurity governance, in *Journal of Cyber Policy*, vol. 6(1), 123–142.
- Roberts A., Choer Moraes H. and Ferguson V. (2019). Toward a geo-economic order in international trade and investment, in *Journal of International Economic Law*, vol. 22, 655–676.

- Sheikh H. (2022). European Digital Sovereignty: A Layered Approach, in *Digital Society*, no. 1, 1-25.
- Sjöstedt G. (1977). *The External Role of the European Community* (Saxon House).
- Soler E. (2024). Cracks in EU Foreign Policy: Exposing Divisions over Palestine and Israel amidst the Gaza War, in *IE Med*, <https://www.iemed.org/publication/cracks-in-eu-foreign-policy-exposing-divisions-over-palestine-and-israel-amidst-the-gaza-war/>.
- Steinbach A. (2023). The EU's turn to 'strategic autonomy': Leeway for policy action and points of conflict, in *European Journal of International Law*, vo. 34(4), 973–1006.
- Vangeli A., Bojinović Fenko A. and Kočan F. (2025). Doomscrollers by Day, Gamers by Night: Patching Digital Sovereignty in the Western Balkan States Amid US-China Technological Rivalry, in *Journal of Contemporary European Studies*, 1–23.
- Vignoli V., Onderco M. and Kalhousová I. (2025). Who cares: Why the Israeli–Palestinian conflict matters (more) to some EU member states, in *Journal of Common Market Studies*, Advance online publication, <https://doi.org/10.1111/jcms.70058>.
- Wodak R. (2019). Entering the 'post-shame era': The rise of illiberal democracy, populism and neo-authoritarianism in Europe, in *Global Discourse*, vol. 9(1), 195–213.
- Wunderlich J.-U. (2011). European Integration, Global governance and international relations, in J.-U. Wunderlich and J. D. Bailey (eds.), *The European Union and Global Governance: A Handbook* (Routledge), 48–55.

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

Lessons from Feminist Foreign Policies: Rethinking the EU's Role in Promoting Peace and Human Rights Amid Anti-Gender Politics and Authoritarian Populism

ESRA AKGEMCI

Associate Professor, Department of International Relations, Selçuk University (Turkey)

✉ esra.akgemci@selcuk.edu.tr

🌐 <https://orcid.org/0000-0003-4119-2443>

ABSTRACT

Gender equality and human rights are core principles of EU policies, backed by the Women, Peace, and Security (WPS) agenda and UN Security Council Resolution 1325. These frameworks laid the foundation for feminist foreign policy (FFP), emphasising gender equality, and many countries have adopted them. However, the growth of anti-gender politics and authoritarian populism threatens progress in gender justice and human rights. Drawing on Judith Butler's theoretical framework, which views anti-gender politics as a psychosocial phenomenon, this paper argues that Butler's concept of the "phantasmatic scene" reveals the psychological logic underlying authoritarian populism. It also examines how FFP can serve as a political tool to confront the phantasm and how the EU can adapt its gender equality policies in response. The paper recommends viewing anti-gender mobilisation as a global security risk, empowering women's and LGBTIQ+ movements in policymaking, and adopting intersectional approaches to peacebuilding. Without using its institutional power to combat anti-gender politics, the EU's dedication to peace, democracy, and equality will remain inadequate.

Keywords: feminist foreign policy, anti-gender politics, authoritarian populism, phantasmatic scene, EU, human rights, peace

ATHENA

Volume 5.2/2025, pp. 240-288

Conference Papers

ISSN 2724-6299 (Online)

<https://doi.org/10.60923/issn.2724-6299/22931>



1. Introduction

Gender equality and human rights are core values of Europe and are central to all EU policies. In 2018, the EU Council called on its member states to fully implement the Women, Peace, and Security (WPS) agenda, which includes UN Security Council Resolution (UNSCR) 1325. Adopted in 2000, this resolution was the first to explicitly recognise the effects of war on women and emphasise the importance of women's roles in conflict resolution and building lasting peace. Fully integrating this resolution into all EU efforts to promote peace, security, human rights, justice, and development has helped ensure the inclusion of gender perspectives throughout EU policies.

UNSCR 1325 was also a groundbreaking and foundational element of feminist foreign policy (FFP), which was first adopted in Sweden in 2014. FFP extends beyond traditional notions of state security by focusing on human security and aiming for a sustainable and peaceful future. The Swedish model is vital for countries pursuing a similar path, such as Canada, Spain, France, Germany, Luxembourg, and Mexico. However, in 2022, Sweden's new conservative government retreated from this policy, asserting that the "feminist" label could be counterproductive. This withdrawal from the birthplace of the FFP is striking, revealing the extent of ultra-conservative mobilisation against the feminist struggle today.

Feminists have historically encountered resistance when pursuing strategic actions toward gender justice and equality. Today, the rise of authoritarian populism has fortified anti-feminist movements across different regions, eroding institutional, legal, and political progress aimed at fighting gender-based violence. The anti-gender movements and campaigns that have appeared over the past decade, especially in Europe, should not simply be seen as a continuation of the anti-feminist backlash from the 1970s, but instead as part of a new political landscape. In fact, opposition to "gender ideology" has become a key element in the growth of authoritarian populism. In today's world, where neoliberalism faces a crisis and people have lost faith in the future, creating a sense of a "common enemy" helps foster "moral panics" and increases societal anxiety. Authoritarian populism is growing worldwide as a national movement that resonates with ordinary people's

lives through conservative values like God, family, homeland, order, and patriarchy, thus reshaping common sense. In this context, “moral panics” emerge as a “crisis of masculinity”; discourses focused on protecting the family and homeland are efforts to defend masculinity. As Judith Butler (2021) states, the defence of gender equality, now often equated with “communism” or “totalitarianism,” has become a key battleground globally. Therefore, there is an urgent need for feminist perspectives that reveal how power relations are gendered in international politics.

This paper aims to connect political trends with foreign policy and examine how FFP can help oppose the anti-gender agenda, which has become a key element of authoritarian populism. The paper argues that EU gender equality policies need to be restructured to address anti-gender politics and authoritarian populism, drawing on lessons learnt from a decade of FFP experience. To do this, I use the theoretical framework of Butler (2025), who views the anti-gender movement as a psychosocial phenomenon, and incorporate Laplanche’s perspective to understand how deep fears and anxieties are socially organised to fuel political passions. I argue that Butler’s “phantasmatic scene” reveals the psychic logic of authoritarian populism, which generates “moral panics” and weaponises fear to restore patriarchal, nationalist authority. Accordingly, the second section will discuss anti-gender politics as a “phantasmatic scene” for authoritarian populism and the third section will consider FFP as an ethical and political vision to counter the phantasm and confront anti-gender politics. The fourth section will propose three central policies for restructuring EU gender policies as a form of critical imagination that would be powerful to oppose the phantasm. It argues for recognising anti-gender mobilisation as a global security threat, empowering women’s and LGBTIQ+ movements in policymaking, and adopting intersectional approaches to peacebuilding. Finally, the fifth section will offer some general conclusions about the potential of the EU’s institutional power in its commitment to peace, democracy, and equality.

2. Anti-Gender Politics as a “Phantasmatic Scene” for Authoritarian Populism

Feminists usually see gender not as a biological fact but as a sociocultural process of becoming a historically rooted, performative, and relational category that affects how people perceive themselves. Simone de Beauvoir’s

famous statement, “*One is not born, but rather becomes, a woman*” (1949), lays the groundwork: gender is not an innate condition but a process of becoming shaped by social and cultural influences. Later feminists highlighted that gender is constructed by norms, institutions, and power relationships, not solely determined by sexed bodies. Since it is often misunderstood or conflated with other categories, much of feminist literature focusses on clarifying what “gender” is not. A common mistake is to view gender studies as only women’s studies. While feminism has traditionally concentrated on women’s oppression, gender is an analytical category that includes the relations, roles, and expectations that shape all genders (women, men, non-binary, etc.). Another common misunderstanding is to confuse gender (social, cultural, symbolic roles and identities) with sex (biological traits). Ann Oakley (1972) is often recognised as the first feminist to systematically separate sex (biologically determined) from gender (socially and culturally constructed). This paved the way for feminist theory to explore how power, norms, and institutions shape gendered identities. Accordingly, Gayle Rubin (1975) theorised the “sex/gender system” as the social structures that transform biological sex into socially meaningful gender, and Joan Scott (1986) recognised gender as a central analytical category in history and the social sciences.

Today, after much discussion, some still believe that “gender” is the same as “women” or “sex.” More often, others see it as a hidden way of referencing “homosexuality” or queer identity. Queer and feminist theorists (e.g., Judith Butler, Eve Kosofsky Sedgwick) have shown how norms of gender and sexuality overlap, but they are not interchangeable. Accordingly, gender teaches us how to appear “properly” male or female; sexuality influences who we are expected to desire. Together, they form the “heteronormativity,” the cultural assumption that heterosexual desire is natural and that sex (male/female), gender (masculine/feminine), and sexuality (heterosexual) must align (Butler, 1990). Adrienne Rich (1980) already argued that social systems force women into heterosexual relations as part of patriarchy. Butler

(1990) and Sedgwick (1985) extended this by showing how “gender performance” is shaped to secure heterosexuality. By exposing that neither gender nor sexuality is natural, but both are cultural scripts, queer and feminist theorists reveal the possibility of resistance and new forms of desire. This growing queer potential has faced one of the most conservative reactions of all time, the so-called “anti-gender ideology” movement, which has grown since the 2010s in Europe, Latin America, and beyond. This movement deliberately distorts the meaning of “gender.”

The “anti-gender ideology movement,” as Butler (2025) states, views “gender” as a single, unified idea that threatens society and consolidates various fears and anxieties under one label. Supporters of this movement argue that when feminists, researchers, or policymakers use the term “gender,” it is really a “Trojan horse” for promoting homosexuality, same-sex marriage, or LGBTIQ+ rights. They claim that gender theory denies “God-given” or “natural” differences between men and women. Italian Prime Minister Georgia Meloni’s warning that “gender ideology” will strip everyone of their sexed identity is a recent example of that. Some others see gender as “the work of the devil,” a modern rival to God that must be eradicated at all costs, and they demonise all those who support gender. By describing gender as socially constructed, they assert feminists seek to eliminate sexual differences. They portray “gender ideology” as undermining the “natural family,” parental authority, and the traditional roles of mother and father. In Russia, it has been labelled a threat to national security, while the Vatican has called it a danger to both civilisation and “man” itself (Butler, 2025). In various parts of the world, school curricula that include gender equality or sexual diversity are often called “indoctrination” or the “corruption of children.” Pope Francis’s description of gender education in schools as “ideological colonisation” and his comparison of it to “Hitler Youth” in 2015 were among the most provocative remarks, setting the tone for his anti-gender campaign (Dempsey, 2020). Butler (2025) notes that

In recent US campaigns to keep “gender” out of the classroom, “gender” is treated as code for pedophilia or a form of indoctrination that teaches young children how to masturbate or become gay. The same argument was made in Jair Bolsonaro’s Brazil because gender calls into question the natural and normative character of heterosexuality, and that once the heterosexual mandate is no longer firm, a flood of sexual perversities, including bestiality and pedophilia, will be unleashed upon the earth. (Butler, 2025).

The anti-gender movement is not new. As Kandiyoti (2022) demonstrates, as early as the 1990s, there was a deliberate effort to discredit the idea of gender as a social construct and to portray gender relations not as human rights issues but as matters of doctrinal necessity. The 1990s also saw “gender” being portrayed as a threat to the family and biblical authority by the Roman Catholic Council for the Family. Since then, it has undergone changes that reflect the Vatican’s shifting political influence. Today, the term “gender ideology” is mostly used as a rallying cry by conservative and right-wing populists to unite various concerns—about women’s rights, reproductive rights, same-sex marriage, trans rights or even EU/UN human rights frameworks (Butler, 2025). Gender issues have increasingly been exploited as tools for demagoguery and populist propaganda. Anti-gender ideologies have become part of xenophobic nationalism, with claims that gender equality platforms are a “product of the liberal West” (Kandiyoti, 2022). In this way, “gender” becomes a broad symbol of moral decline, globalisation, or liberal elites (Butler, 2025). Russia’s President Vladimir Putin’s referring to Europe as “Gayropa,” claiming that gender is a Western construct that will destroy the concepts of mother and father, is a clear example of this.

In her recent book “Who’s Afraid of Gender?” (2025), Butler borrows the term “phantasmatic scene” from psychoanalyst Jean Laplanche to describe how the anti-gender movement shapes fear and anxiety into a distorted but compelling worldview. Butler emphasises that this is not a private daydream

but a shared psychosocial setup of fears, projections, and reversals. It combines unconscious anxieties with social narratives. In this scene, “gender” is seen as a destructive force—corrupting children, breaking up families, and weakening the nation. It functions as a placeholder that condenses many fears—about sexuality, religion, race, and the nation—into one powerful symbol. Hatred is justified through moral righteousness. The anti-gender movement portrays LGBTIQ+ and feminist communities as the true destroyers. For Butler (2025), calling anti-gender discourse a phantasmatic scene shows it is less based on rational argument and more on a hallucinatory politics of fear. Exposing this involves not just fact-checking, but also dismantling the fantasy structure and creating counter-imaginaries that support freedom, equality, and sustainable lives. Butler uses “phantasmatic scene” to show how gender serves as a battleground for collective fantasies of destruction, where fear and moral panic are projected onto marginalised groups—serving as a way of organising hatred that seems real but is ultimately false.

Butler (2025) emphasises that the weaponisation of this frightening illusion of “gender” is fundamentally authoritarian and can be exploited by those seeking to strengthen state power and restore a “secure” patriarchal order. This aspect is crucial to understanding anti-gender movements not only as a backlash against progressive movements, but also as a political project to reestablish a patriarchal system. I argue that Butler’s conceptualisation of anti-gender discourse as a phantasmatic scene helps us understand how authoritarian populists legitimise authoritarian responses today. Bringing Butler’s concept of the phantasmatic scene into dialogue with Stuart Hall’s insights on moral panics will reveal that authoritarian populists do not just “discover demons”; they stage a phantasmatic scene where the “enemy” (feminists, LGBTIQ+ people, immigrants, “globalists”) threatens the very survival of the family, nation or faith.

Stuart Hall (1978) has demonstrated that during moments of crisis, popular anxieties and perceived threats to the state merge through mechanisms such

as the “discovery of demons,” the “identification of folk devils,” and the “mounting of moral campaigns.” These processes generate waves of moral panic and help justify authoritarian efforts (Hall et al. 1978; Hall 1988). Butler’s idea of the phantasmatic scene enriches this framework by emphasising the psychic side of such politics: crises are not only described, but also staged as hallucinatory scenarios where broad fears are projected onto a single symbolic enemy. In these scenes, “gender” or “the left” often appear as destructive forces that corrupt children, dissolve families or weaken national unity. By reducing complex insecurities to an imagined existential threat, authoritarian populists stoke paranoia and mobilise collective emotions. Together, Hall’s concept of moral panic and Butler’s idea of the phantasmatic scene reveal how authoritarian populists create existential fears that justify a return to a “secure” patriarchal order—where fixed gender roles and authoritarian leadership are presented as defences against chaos.

Hall first introduced the idea of authoritarian populism to analyse Thatcherism, illustrating how her government manipulated moral panics through the “discovery of demons” and the “identification of folk devils.” In Thatcher’s Britain, these folk devils were not only racialised youth accused of “mugging,” but also feminists, queer activists, and trade unionists— all depicted as threats to the moral and social order (Hall et al. 1978; Hall 1988). Thatcher herself was openly hostile toward the LGBTIQ+ community, implementing policies like Section 28, which banned “promotion of homosexuality.”¹ Hall’s explanation highlights the political mechanics of fear. Still, Butler’s concept of the phantasmatic scene adds an important layer: authoritarian populism functions not simply by naming enemies, but by creating hallucinatory scenes where broad anxieties are projected onto vulnerable groups. Butler explains how “gender” becomes a focal point for fears about cultural decline, sexual deviance, and national collapse. This framework is crucial for understanding why authoritarian populism resonates

¹ See Parker (2022) for how Jeannette Winterson’s novels expose Thatcherism as a heteropatriarchy.

with people today. Leaders like Jair Bolsonaro and Donald Trump use similar tactics: they link feminism, LGBTIQ+ rights, and progressive education to threats against children, family, and the nation, creating existential dangers that call for authoritarian responses. Bolsonaro's attacks on "gender ideology" and Trump's demonisation of "critical race theory" and "trans ideology" serve as emotional displays of crisis — charged scenes that stir paranoia and promise salvation through patriarchal revival. Seen this way, Hall explains the political strategy (moral panic as a tool for securing consent), while Butler illuminates the psychic logic (phantasmatic projections that transform political disagreement into existential fear). Together, their insights reveal how authoritarian populists from Thatcher to Trump and Bolsonaro weaponise cultural anxieties, casting LGBTIQ+ people, feminists, and migrants as the enemies of "the people" to legitimise authoritarian rule and a return to a "secure" patriarchal order.

Jeremiah Morelock and Felipe Ziotti Narita (2021) identify the social and narrative trends that develop as the realms of authoritarianism and populism blend in the United States and Brazil. A key dynamic alongside the creation of charismatic leaders is the mythological use of the past, seen in phrases like "Make America/Brazil great again." Mythological references to the past inspire "retrotopia"—a term defined by Zygmunt Bauman (2017) as a utopian vision focused on an abandoned past—especially when there are no promising future utopias. This can be politically appealing because it fills or replaces "the loss of sense of the masses of people" who are disillusioned by the uncertainties built into the current social system, which they once regarded as more stable and secure (Akgemci, 2022). Similarly, Butler (2025) states that promoting a desire to restore masculine privilege seeks to craft an "idealised past" whose revival aims to target sexual and gender minorities and undo progressive policies and rights. The backlash we see against "gender," according to Butler, is part of this larger effort to restore authority that aims to legitimise authoritarian regimes as legitimate forms of paternalism, fulfilling this dream. Butler (2025) observes:

The mobilisation of anti-gender sentiment by the Right relies on the credibility of this past dream for those vulnerable to authoritarianism's temptations. In this way, the fears are neither entirely manufactured nor completely real, as they already exist. This idea of the past belongs to a fantasy whose syntax rearranges elements of reality to serve a driving force that makes its own operation opaque. The dream functions only as a phantasmatic organisation of reality, offering a range of examples and accusations to support the political case it aims to make. Stripping people of rights in the name of morality, the nation, or a patriarchal fantasy is part of a broader logic amplified by authoritarian nationalism (Butler, 2025).

Butler (2025) also highlights that this project is fragile because the patriarchal order it seeks to restore never fully existed in the way it is now being presented. By showing that the references to the past — especially the patriarchal order — are themselves illusions, Butler (2025) deepens the critique of authoritarian populism. Just as Butler notes that the patriarchal order the anti-gender movement longs for is an impossible restoration, authoritarian populism also remains fragile: it relies on constantly fostering fear and fantasy, but its utopian past is always out of reach. It depends on establishing legitimacy by returning to a past that never truly existed. Like Butler's critique of patriarchy, its "retrotopian" promise is a false vision — persuasive but unsustainable. This gap between promise and reality is a vulnerability that can be politically challenged.

Similarly, Schleusener (2020) introduces the idea of "retrotopian desire," drawing on Bauman's concept of "retrotopia," to explain the longing for an idealised past that fuels right-wing populist politics. It also explores how right-wing populism uses nostalgia and gendered imaginaries to respond to neoliberal crises. Accordingly, right-wing populist movements frame a

counterrevolutionary project: reclaiming “lost” social life rooted in the heteronormative family, the church, the homeland, and patriarchal gender relations. These narratives target both “inner enemies” (feminists, LGBTIQ+ activists, liberal elites) and “outer enemies” (migrants, cosmopolitan outsiders), framing gender as a key battleground in political conflict. Schleusener (2020) links this to class politics, noting that neoliberal restructuring has destabilised traditional male breadwinner roles and fuelled resentment. Authoritarian populism channels these anxieties into gendered, nationalist and nostalgic imaginaries. In summary, retrotopian desire acts as the cultural and emotional core of authoritarian populism, uniting economic fears, gendered masculinity crises and nationalist hopes.

In addition to the most notable cases, such as Donald Trump (2017-2021; 2025-) in the U.S. and Jair Bolsonaro in Brazil (2019-2022), during the past decade, authoritarian populism has gained momentum worldwide with the rise of right-wing and far-right leaders, including Viktor Mihály Orbán (2010-) in Hungary, Narendra Modi (2014-) in India, Andrzej Duda (2015-2025) in Poland, Rodrigo Duterte in the Philippines (2016-), and Andrej Babiš (2017-2021) in the Czechia. Authoritarian populism also appears in Shinzo Abe’s 2012 campaign with the slogan “Take Back Japan,” which promotes a national revival in Japan, Recep Tayyip Erdoğan’s repressive response to the failed July 2016 coup attempt in Turkey, and Marine Le Pen’s 2017 campaign to push nativist policies against Muslims and non-European immigrants in France (Gokay, Xypolia, and Mandelbaum, 2017, 3). Recent cases confirm that authoritarian populism is not a fading trend but a growing global phenomenon. In El Salvador, Nayib Bukele has centralised power in the presidency, weakened checks and balances, and redefined democracy through his “mano dura” security strategy. In Europe, Giorgia Meloni became Italy’s first far-right prime minister in 2022, merging nationalism with socially conservative values, while Robert Fico’s return to power in Slovakia has echoed Orbán’s illiberal model. Meanwhile, Spain’s far-right party Vox has pushed anti-immigration and anti-gender narratives into the mainstream,

reshaping national debates. Furthermore, the outcome of the Brexit referendum in 2016 showed that even if authoritarian populist trends do not dominate national politics, they can still influence the policy agenda by promoting anti-EU and anti-immigrant attitudes (Norris and Inglehart, 2019, 12). Therefore, it can be argued that authoritarian populist sentiments are expressed not only by right-wing leaders and political parties, but also in public opinion and social movements.

With the outbreak of the European migrant crisis in 2015 and the Central American migration crisis in 2018, xenophobic nationalism drew voters of all ages toward right-wing extremes and laid the groundwork for right-wing populism to shape the current “authoritarian turn.” However, one could argue that authoritarian populism can also be combined with economic protectionism or left-wing social policies. Still, I maintain that the phenomenon emerged in response to specific conditions, where far-right leaders used populist rhetoric to appeal to the “people” and advocate for authoritarian measures. In other words, a leftist leader can also be both authoritarian and populist; however, the term “authoritarian populism,” coined by Hall, describes a distinct form of far-right populism. It is crucial to understand how authoritarian values intersect with right-wing populist rhetoric in today’s context. Scholars have identified several factors—such as economic insecurity, financial globalisation, technological change, mass immigration, and the failure of representative politics—to explain how right-wing populism has gained so much power today. The authoritarian element stemmed from the harsh repression of leaders, fuelled by growing anxiety and hopelessness in society. A significant work also focused on discovering the relationship between far-right populism and anti-gender politics, especially in Europe.

Populist radical-right movements and parties rally around what Dietze and Roth (2020) call an “obsession with gender”: demonising feminism, LGBTIQ+ rights, and “gender ideology” while endorsing heteronormativity, patriarchal family models, and masculinist identity politics. These highlight

gender as a key analytical tool for understanding the success and ideological strength of right-wing populism, rather than treating it as a secondary or cultural side issue. According to Roman Kuhar (2023), the anti-gender movement, which includes radical nationalist parties and right-wing populists, has become a potent transnational force because it redefines feminist, LGBTIQ+, and equality struggles as a threatening “gender ideology,” offering simple and fear-based narratives that resonate with various societal crises. Even in countries like Poland and Hungary, anti-gender ideology has become the official stance of the ruling political elites. The “traditional family” is where the anti-gender movement and the radical right intersect in this framework, being presented as the main thing to be protected from “abnormal” LGBTIQ+ lifestyles or radical feminism (Kuhar, 2023, 120). Considering “gender ideology” as a tactic that intersects with debates within the Catholic Church and the recent rise of right-wing populism in Europe, Kuhar and Paternotte (2017) illustrate how both anti-gender and populist groups operate on a binary opposition: “the pure people” vs. “the corrupt elites.” Here, gender functions as a “symbolic glue” that unites diverse actors (religious, secular, far-right, conservative) against a common enemy. Kuhar and Paternotte (2017, 15) describe “gender ideology” as an empty signifier, enabling coalition building with various actors precisely because of its “populist emptiness.” Populists blame international and supranational authorities, often called “Brussels” in the European context, for secretly imposing “gender ideology” on ordinary people (Kuhar and Paternotte, 2017, 14). Especially in Central and Eastern Europe, referenda serve as a key strategy. Anti-gender campaigns in Croatia, Slovenia, Slovakia, and France utilised referenda as “the people’s voice” to oppose elites.

Similarly, Graff and Korolczuk (2022) see the current wave of anti-gender mobilisation not just as a continuation of an older conservative backlash, but as a new ideological and political movement closely linked to right-wing populism. They highlight that anti-gender campaigns and right-wing populism support each other through an “opportunistic synergy”: populist

parties use anti-gender rhetoric to stir emotions and depict themselves as defenders of “the people” against liberal elites, while anti-gender groups gain strength and resources from populist alliances. In this context, understanding how gender plays a vital role in the populist moment is crucial. Graff and Korolczuk (2022) argue that conflicts over gender equality are not minor “cultural issues,” but essential arenas that shape the future of democracy. Similarly, Kourou (2022) states that right-wing populism and anti-gender movements are “two faces of the same coin”: both leverage social discontent, mobilise resentment against elites, and define “the people” in opposition to feminists, sexual minorities, and liberal institutions. Their cooperation centres on gender in populist politics, undermines democratic norms, and turns anti-gender ideology into government policy where populists hold power (e.g., Poland, Hungary). This convergence poses a serious challenge to gender equality and democratic values globally. Fassin (2020) contributes to this debate by arguing that anti-gender politics should be understood as part of the global rise of illiberal neoliberalism, where neoconservative and neoliberal logics converge. Instead of being a contradiction, Fassin (2020) sees their alliance as functional: neoliberalism shifts responsibility from the state to the family, while neoconservatism enforces patriarchal norms to support this privatisation. Therefore, anti-gender campaigns are not just cultural conflicts but key tools of contemporary populism, transforming socioeconomic dissatisfaction into moral panic about “gender” and fuelling authoritarian politics across Europe and Latin America.

Lastly, Mehring and Wojnicka (2025) make a significant contribution to the literature on right-wing populism and gender by combining both qualitative and quantitative research across 16 European countries, providing a cross-national, comparative perspective. They introduce the concept of “protective masculinity”—defined as the intersection of gender-unequal and nativist attitudes—into the study of populist radical right (PRR) voting. This shifts the focus from a simple male/female binary to masculinity as a social construct that influences political behaviour. This is a clear improvement over

earlier explanations that focused only on socioeconomic status, political efficacy or anti-immigrant attitudes. Mehring and Wojnicka (2025) argue that masculinity — especially protective masculinity — is a crucial yet underexplored element in understanding PRR support and the gender gap. PRR parties use protective masculinity by portraying themselves as defenders of women and the nation against “foreign” masculinities, especially Muslim men. Although this only partially explains why men are more likely to vote for PRR, it shows that gender norms and masculinities are central in shaping today’s far-right politics.

Overall, research on anti-gender politics and populism also confirms that the anti-gender phenomenon is best understood as a “phantasmatic staging” of crisis. In this sense, gender acts as a projection surface where authoritarian populists express fears of social collapse, demographic decline, and cultural decay. The success of anti-gender campaigns comes from their ability to simplify complex issues — such as neoliberal insecurity, weakening welfare systems, and changing ideas of masculinity — into emotionally charged displays of danger. Butler’s concept of the phantasmatic scene helps explain the mental reasoning behind this politics: the more phantasmatic the “gender threat” seems, the more convincing it becomes as a justification for authoritarian measures. In conclusion, these works share a common insight: right-wing populism (therefore authoritarian populism) is driven less by rational policy and more by fantasies of restoring lost order and fear, often fuelled by demonising gender equality and sexual rights. As a result, anti-gender politics serve not only as ideology or discourse, but as phantasmatic scenes — hallucinatory yet socially real scenarios that shape desire and hatred, making authoritarian rule seem like a paternalistic “return” to order. That is why it is crucial to envision ethical and political futures to counter authoritarian populisms rooted in phantasmatic scenes that are becoming increasingly ingrained.

3. FFP as an Ethical and Political Vision to Counter the Phantasm and Confront Anti-Gender Politics

Feminist foreign policy (FFP) has been adopted by numerous countries and organisations, enabling them to prioritise gender justice and equality in their foreign policy strategies. However, the adoption of the FFP in Sweden occurred during a global rise in right-wing populism, which led to far-right parties forming governments in several nations. As explained above, authoritarian populism, as a specific form of far-right populism, is rooted in distrust of what is referred to as “gender ideology” and is driven by a reaction against feminist values in global politics. Therefore, successfully launching and implementing new foreign policy initiatives, including feminist foreign policies, requires the ability to challenge anti-gender sentiments (Aggestam, Bergman Rosamond, and Hedling, 2024, 20). Furthermore, FFP must be reconsidered in the context of rising anti-gender politics and be crafted as a strategy to counter it.

Confronting anti-gender politics was central to Sweden’s initial adoption of FFP in 2014. Margot Wallström, who was the Foreign Minister of Sweden back then, referred to “unsettled times” in which “Sweden will take global responsibility by being a strong voice in the world. For freedom, peace, and human rights. For democracy, equality, and solidarity.” In this, “gender equality and a feminist foreign policy” were presented as “building blocks for a foreign and security policy” guided by “the necessity of common security” (Government of Sweden, 2015). Wallström’s successor as Foreign Minister, Ann Linde, also emphasised that:

The rights of women and girls are under attack. Conservative forces are trying to restrict the right of women and girls to decide over their own bodies and lives. Issues relating to women, peace and security, as well as women’s sexual and reproductive health and rights, are especially important to stand up for. To reverse this trend,

courageous action is needed at all levels. This is why we are pursuing a feminist foreign policy. (Government of Sweden, 2020).

The adoption and spread of FFP by other countries can be seen as a success story. Sweden, a nation where gender equality has long been a key part of its foreign policy, has developed and implemented the concept of FFP for nearly ten years, becoming a model others can follow. Wallström introduced the “three Rs,” encouraging Swedish foreign policy actors to adopt policies that promote equal *rights*, fair *representation*, and an equitable distribution of *resources* among men and women, as well as boys and girls. This policy was a part of broader global efforts to promote gender equality internationally, which evolved over the past few decades following the adoption of United Nations Security Council Resolution 1325 (Aggestam and Bergman-Rosamond, 2016, 323). The resolution marked a significant milestone in feminist international relations and global security governance in 2000, establishing the foundation of the Women, Peace, and Security (WPS) agenda and serving as a normative framework for Swedish foreign policy. In 2019, the need for a fourth “R,” *Reality*, was emphasised in Sweden’s Handbook, encouraging the Foreign Service to understand the context in which they are working and engage with local actors, aiming to contribute to strategic, effective feminist foreign policy (Thompson, Ahmed, and Khokhar, 2021, 2). As a result of the FFP, efforts to promote gender equality increased significantly, and Swedish trade policy experienced its most notable change. Before 2014, agencies, missions and embassy sections involved in trade paid no attention to gender issues at all, and they had to start figuring out what FFP meant for Swedish trade policy (Towns, Jezierska, and Bjarnegård, 2024, 1267).

Scholars specifically criticised the role of Sweden as one of the world’s leading arms exporters because the arms trade has come into conflict with the FFP. In 2015, Sweden ended its military cooperation with Saudi Arabia due to concerns about Swedish-made arms being used in the war in Yemen (Foster and Markham, 2024, 58). Still, Sweden was the 13th largest arms exporter in

the world in 2020, posing a significant challenge for FFP, since we see that gender-based violence is increasing in current conflicts and weapons enable this violence. Some feminist scholars have also criticised the Swedish application of FFP for not being transformational enough. They argue that, in fact, these policies do not reshape government structures, reduce militarism or address power imbalances (Foster and Markham, 2024, 58).

Despite its limitations, the launch of the FFP in Sweden sparked a noticeable rise in gender equality initiatives in other regions. Canada launched the world's first Feminist International Assistance Policy (FIAP) in June 2017, further solidifying its leadership in promoting the FFP. Canadian vision for international aid was outlined as follows:

Canada is adopting a Feminist International Assistance Policy that seeks to eradicate poverty and build a more peaceful, more inclusive, and more prosperous world. Canada firmly believes that promoting gender equality and empowering women and girls is the most effective approach to achieving this goal. (Government of Canada, 2018).

For several years after the launch of the FIAP, Canada's explicitly "feminist" approach to foreign policy was limited to its international aid policy. In 2020, Global Affairs Canada released a discussion paper on developing the FFP, acknowledging pressure from civil society and scholars to go beyond aid. The document suggested expanding the feminist approach to trade, diplomacy, migration, environment, and security — but Canada had not yet officially adopted a comprehensive FFP (Thompson, Ahmed, and Khokhar, 2021, 6). Canada offers a good example, showing us that we should not focus only on the content of feminist foreign policy, but also consider the policy ecosystem that determines whether feminist commitments can be practically implemented. Gloria Novović (2024) argues that Canada's FFP remains more rhetorical than substantive because the government lacks the systemic, organisational and individual policy capacity to implement it

effectively. According to Novović (2024), without stronger political will, inter-ministerial coordination and sustained institutional funding, Canada's FFP risks becoming symbolic branding and technical rhetoric rather than a transformative feminist global engagement.

Luxembourg first included the goal of adopting FFP in its Coalition Agreement in 2018 and created an important model often called the “3 Ds”: defence, diplomacy and development. Accordingly, three main priorities were: (1) protecting and promoting the human rights of women and girls; (2) increasing women's representation and participation in multilateral forums, as well as involvement in civil and electoral observation missions; and (3) promoting gender equality within the structures of the Ministry of Foreign and European Affairs (MFEA) (Thompson, Ahmed, and Khokhar, 2021, 8). Luxembourg adopted a National Action Plan (NAP) on Women, Peace, and Security for 2018–2023, linking it to its external efforts in diplomacy, defence, and development, while also strengthening domestic prevention and protection mechanisms. In 2025, it launched a second NAP for Women, Peace & Security (2025–2030), organised around four pillars: (i) participation, (ii) protection, (iii) prevention, relief, and recovery, and (iv) promotion (Government of Luxembourg, 2025). In the second five-year plan, the rollback of women's rights was seen as a global challenge alongside climate change and the emergence of new and rapidly evolving technologies. The plan was adopted to address these challenges, emphasising the importance of full, equal, and meaningful participation of women in peace, security, prevention and conflict management processes, while focusing on issues such as disarmament, non-proliferation, the elimination of sexual and gender-based violence, the fight against impunity and the protection of individual rights (Government of Luxembourg, 2025).

France was among the first countries to adopt the FFP approach in 2019, becoming the fourth country after Sweden, Canada, and Luxembourg to do so. French officials have consistently referred to the French approach as France's “feminist diplomacy,” and after the March 8th op-ed, as FFP

(Thompson, Ahmed, and Khokhar, 2021, 9). Like Canada, France prioritised gender equality during its G7 presidency, focusing on access to education for girls and women, combating gender and sexual violence and elevating the status of African women (Foster and Markham, 2024, 67). The FFP of France is based on the principle that women's and girls' rights, together with gender equality, must be at the heart of its international and European efforts. All areas, such as diplomacy, development, trade, humanitarian work, digital policy, environment, security, among others, are intended to incorporate gender equality. In March 2025, France introduced its International Strategy for a Feminist Foreign Policy (2025–2030), which built on and broadened previous gender initiatives (Government of France, 2025). This new approach stems from an inclusive process that involves ministries, agencies, civil society, and external partners with more than 200 participants across various working groups. Despite strong ambitions, the French FFP faces significant challenges. The High Council for Gender Equality (*Haut Conseil à l'Égalité*) released its 2023 accountability report, highlighting key shortcomings: the policy lacked an inclusive conceptual definition, political support was weak, and resources were limited. Critics also argue that France's FFP sometimes lacks clarity in its implementation, with limited transparency around budgets specifically allocated for gender equality. Some observers emphasise colonial legacies and power imbalances, noting that as a former colonial power with close ties to Africa, the feminist diplomacy of France must address critiques of neocolonialism or paternalism in its foreign actions (Foster and Markham, 2024, 68).

In 2020, Mexico became the first country in the Global South to adopt FFP. This policy is based on five principles: (i) implementing policies to promote gender equality and a feminist agenda; (ii) achieving gender parity at all levels within the foreign ministry; (iii) combating all forms of gender-based violence, including within the foreign ministry; (iv) making equality visible; and (v) practising intersectional feminism (Government of Mexico, 2020). To put this policy into practice, specific steps were outlined for each

area, such as the presentation of the Manual of Foreign Policy Principles, the Foreign Ministry support for the *HeforShe* program, certifications of labour equality and non-discrimination, the development of training sessions, workshops, and working groups with key actors, and establishing a safe, violence-free zone near the Foreign Ministry. Mexico also linked its FFP to the implementation of the 2030 Agenda for Sustainable Development and advocated in the fight against intersectional injustices, as seen at COP27. The main critique of Mexico's ambitious FFP agenda was the gap between the country's goals and the current state of gender relations. Eradicating gender-based violence was a top priority for implementing the FFP since Mexico has one of the highest femicide rates in the world. However, funding was cut for *Inmujeres*, the Mexican federal agency that coordinates gender equality policies and fights violence against women, at the same time as the FFP was announced (Foster and Markham, 2024, 70). Women's rights activists criticised government inaction and the lack of acknowledgement of domestic violence, state violence, and femicide (Deslandes, 2020). Another issue that makes FFP problematic is the rapid increase in militarisation under President Andrés Manuel López Obrador (AMLO). Easier access to firearms, a 50 percent increase in the armed forces budget, and a 25 percent rise in military personnel, along with the military taking over some public security tasks, contradict feminist principles. AMLO has also faced criticism for his government being based on a "masculine vision" (Foster and Markham, 2024, 70). Following the disappointment due to his hostility towards the feminist movement, there have been expectations for his successor, Claudia Sheinbaum, Mexico's first female president, to strengthen the FFP. She is seen as re-engaging Mexico on the global feminist diplomacy stage with the potential to align foreign policy with domestic feminist agendas.

Spain officially adopted FFP and became the sixth country to do so in 2021, releasing its Guide to Spain's Feminist Foreign Policy as a plan to promote gender equality in its international efforts. The government describes it as part of its commitment to the 2030 Agenda and Sustainable Development

Goal 5 (gender equality). Spain's FFP is organised around five guiding principles: (i) Transformative approach – moving beyond a symbolic adoption to change institutional culture and practices within the foreign service; (ii) Committed leadership – ensuring those in leadership roles take responsibility for promoting gender equality; (iii) Ownership – integrating gender equality into management, resource allocation, and decision-making across the foreign ministry; (iv) Inclusive participation and alliances – working with other ministries, civil society, research institutions, and international networks; and (v) Diversity and intersectionality – acknowledging overlapping forms of discrimination and including them in analysis and programs (Government of Spain, 2021). In terms of lines of action, Spain combines: (i) Gender mainstreaming: embedding gender perspectives into all phases of foreign policy; (ii) Bilateral and regional diplomacy: ensuring gender issues appear on agendas of official visits, negotiations, and diplomatic dialogues; promoting women's organisations in partner countries; (iii) Multilateral diplomacy: promoting gender equality in UN forums, EU external action, conventions, and global agreements; and (iv) Monitoring and resources: creating mechanisms to track implementation, and an Equality Unit in the Foreign Ministry (Government of Spain, 2021). The guide details instruments, stakeholders, monitoring efforts, and specific actions, making the FFP more operational rather than merely declarative. It employs a dual approach: strengthening priority areas in foreign service while incorporating gender throughout all external policies. Its focus on structural change and intersectionality highlights its significance as a key case study.

The German government included the FFP in its coalition treaty in 2021. After initially mentioning it in its coalition agreement, the German Federal Foreign Office further developed its FFP with guidelines published in March 2023 (German Federal Foreign Office, 2023). Building on the rights, representation, and resources framework initially laid out by Sweden, the guidance explains how the German foreign service should foster an “internal culture that is free of discrimination, one that values our employees' diversity,

nurtures it, and harnesses its potential” (Foster and Markham, 2024, 72). Critics argue that Germany’s guidelines on FFP are still a “work in progress,” and mechanisms for transparency and accountability in foreign policymaking need to be strengthened. Self-reflexivity is seen as a key feminist principle in the guidelines, paving the way for a more transformative foreign policy (Hauschild and Stamm, 2024, 3-4). However, the key challenge to such a policy was the war in Ukraine, which sped up debates about what a feminist peace and security policy should look like. A new National Security Strategy (NSS) was also established alongside FFP in the coalition treaty in 2021. Pragmatists in government argued that FFP is compatible with arms deliveries for defence, while normative advocates opposed weapons, viewing militarisation as a patriarchal practice (Pierobon, 2024, 286). Tensions persist between feminist ideals and strategic military development, and its success relies on whether resources, representation and accountability measures are effectively implemented.

Chile became the first South American country to establish a formal FFP framework. The new FFP was officially announced when former student leader and leftist President Gabriel Boric took office on March 11, 2022, making the announcement part of a new government. The idea of FFP was already part of the 2021 electoral campaign and the new government successfully incorporated it into Chile’s traditional foreign policy principles (Thomson and Wehner, 2025, 14). The policy aims to establish gender equality and non-discrimination as core principles in Chile’s foreign affairs, promoting respect, protection and the full and equal enjoyment of all human rights for LGBTIQ+ individuals. It also seeks to embed gender mainstreaming into diplomatic, consular, multilateral, and bilateral activities (Government of Chile, 2022). The Sub-secretariat of Foreign Affairs is responsible for leading the design, monitoring, and implementation of the policy, including establishment of a follow-up mechanism. However, some scholars warn that Chile’s FFP 2024–2025 action plan may fall victim to “purple washing”—that is, using feminist language symbolically to boost

reputation without enacting meaningful structural change. Thomson and Wehner (2025) argue that the Chilean FFP largely emerged from favourable domestic mobilisations, political entrepreneurs, and Boric's left-leaning government. They emphasise that this adoption remains very fragile and dependent on various factors. The critique is that Chile's FFP is politically opportunistic, lacks strong institutional support and is at risk of becoming more symbolic than meaningful if it doesn't outlast the current government. Still, Chile's adoption helps expand FFP beyond countries in the Global North. It provides South America with a local example, alongside Colombia, which is a relatively recent case that is still in the early stages of FFP implementation.

In November 2022, the Dutch government aligned itself with a growing group of countries, including Germany, Spain, France, Canada, Mexico, and others. It announced its plan to implement FFP, emphasising equality and equal rights across all aspects of Dutch foreign policy. According to the official policy paper, "Feminist foreign policy means protecting human rights and promoting meaningful participation in decision-making by women and LGBTIQ+ people. The focuses of feminist foreign policy are rights, representation, resources, and reality check" (Government of the Netherlands, 2022). However, after an evolution report found that gender in the context of FFP is mainly seen as synonymous with women and that gender mainstreaming does not sufficiently address more fluid gender identities, the Dutch government, as of June 2022, committed to extensive consultations with civil society to ensure the meaningful development of this policy (Foster and Markham, 2024, 73). The government released a Feminist Foreign Policy Handbook in November 2024 to provide practical guidance for implementation. The handbook outlines seven priority areas: embedding gender perspectives into policy cycles; gender-sensitive budgeting; addressing root causes of inequality; inclusive consultation with civil society; monitoring and evaluation of feminist policies; institutional strengthening; and organisational change (within the ministry and related bodies). Although

these steps are part of a transformative agenda, critics argue the Dutch FFP needs to more strongly address colonial legacies, global power imbalances, and various intersecting forms of discrimination (race, class, disability) instead of using a one-size-fits-all approach.

Meanwhile, after serving as a model for many states, the Swedish government revoked its FFP on the same day the newly elected liberal-right coalition government took office in October 2022. The new Foreign Minister, Tobias Billström, announced that the Swedish FFP would end and explained: “The use of the label feminist foreign policy has obscured the contents of our policy. This is why the government will discard its use. But we will always support gender equality” (Towns, Jezierska and Bjarnegård, 2024, 1263). This was symbolically important because it marked the world’s first withdrawal of an FFP, raising fears of a wider international backlash. However, Towns, Jezierska, and Bjarnegård (2024) argue that due to international norms, decentralised implementation and role expectations, its practices and influence are difficult to completely reverse. The FFPs, once institutionalised, are more resilient than opponents expect. The discontinuation of FFP also signifies a major shift in Sweden’s current foreign and security policy, prompted by Russia’s full-scale invasion of Ukraine and Sweden’s subsequent application for NATO membership, which shifted the focus of foreign policy primarily toward national security interests in the region. (Aggestam, Bergman Rosamond, and Hedling, 2024, 96). This suggests that the worsening security environment carries a risk of moving away from normative commitments and values.

What lessons can we learn from these cases? Three stand out. First, although anti-gender politics have become a significant threat to democratic governance and human rights, most feminist foreign policies lack strong strategies and tools to directly confront this backlash. Second, countries implementing FFP often fail to develop lasting institutional capacity and resources to support and empower women’s and LGBTIQ+ movements as meaningful voices in policymaking, resulting in limited or symbolic civil

society participation. Third, while intersectionality is often cited as a guiding principle, in practice it frequently remains a rhetorical statement rather than a consistently used framework, resulting in fragmented and uneven implementation across ministries and policy areas. Having principles is one thing; coordinating budgets, institutional mandates, staff capacity, evaluation frameworks, and sustained political commitment is another. Observers are watching to see whether the FFP remains rhetorical or becomes substantive. FFP requires coordination across trade, defence, climate, migration, and justice, which can create institutional tension — a significant obstacle to the FFP, which promises a “whole-of-government” transformation. This institutional tension can be managed by combining strong political leadership, cross-ministerial structures, effective accountability mechanisms, cultural change, and engagement with civil society. Without these, FFP risks being siloed in development or diplomacy while other ministries continue with business as usual. Furthermore, managing institutional tension in FFP requires more than just coordination mechanisms; it also needs significant institutional transformation. High-level mandates, inter-ministerial structures, accountability frameworks, and engagement with civil society are essential steps. However, for these measures to be effective and sustainable, gender equality must become a firmly rooted norm within government institutions. When gender equality is integrated as a shared value — rather than a fluctuating or contested principle — ministries are more likely to align their policies rather than resist them. In this way, institutionalising gender equality as a lasting norm works hand in hand with structural reforms, reducing friction between sectors such as trade, defence, climate, and migration, and enabling FFP to serve as a genuinely cross-cutting agenda rather than a siloed initiative.

Thomson (2024, 57) considers gender equality in foreign policy a “fluctuating norm” that appears not only in the actions of pro-gender equality countries, but also in its challenge by anti-gender actors and its adoption in illiberal states. The concepts of “gender bashing” and “gender washing”

represent two opposing political strategies regarding gender equality norms. “Gender bashing” involves openly rejecting and demonising gender equality, often framing it as a foreign or corrupting “gender ideology” that threatens the family, nation, or religion. In contrast, “gender washing” refers to a superficial adoption of gender equality language and symbols without meaningful policy change — a form of superficial compliance that aims for legitimacy while maintaining structural inequalities. That is why Thomson (2024, 61) describes gender equality as a global norm that is viewed as “content-in-motion” and in a state of “flux” compared to other established norms. It is essential to understand the contrasting ways in which gender equality functions as a normative force in foreign policy. FFP, then, must be designed to oppose “gender bashing,” which seeks to mobilise conservative supporters and legitimise authoritarian populist leadership, and against “gender washing,” which exploits women’s rights discourse to gain international prestige.

The adoption of gender equality from a temporary “fluctuating norm” to a lasting and deeply rooted standard in international politics is critically essential. While increased support for gender equality by liberal democratic countries has helped promote the norm, its long-term strength cannot rely solely on national backing. For gender equality to transition from a fragile, disputed idea to a firmly institutionalised one, it must also be integrated into powerful international and regional structures that actively defend human rights. The EU offers a strong example: as a cross-border entity with both normative influence and policy tools, the EU can protect gender equality even when facing domestic opposition within its member states. By incorporating gender equality into its peace, security, and human rights efforts, and by making it a core principle in both its external and internal policies, the EU can help transform gender equality from a vulnerable, shifting goal into a constant element of democratic governance and global order.

To oppose anti-gender ideology politics, as Butler (2025) recalls, we need transnational coalitions that gather and mobilise everyone they have targeted.

Rethinking how the EU can adjust its gender policies to build such a transnational coalition is a worthwhile task. We need to develop a sense of solidarity and a strong ethical and political vision capable of exposing and defeating the brutal norms promoted under the banner of the anti-gender ideology movement (Butler, 2025). I argue that by reshaping the FFP as a tool against anti-gender politics, foreign policies can become a powerful means for promoting gender equality and global peace. Lessons from a decade of experiences in countries implementing FFP can help the EU craft an ethical and political vision to oppose anti-gender politics, which has become a “phantasmatic scene” for authoritarian populism.

4. EU to Counter Anti-Gender Politics: Three Pillars Against the Phantasm

Hence, what role should the EU play in this complex landscape? Before urging the EU to act on this section, it is helpful to look at examples from Europe and elsewhere that show how anti-gender politics is becoming an increasing threat under authoritarian populists, as discussed earlier.

In Poland, anti-gender politics under Andrzej Duda (2015-2025) have functioned as a classic phantasmatic scene: a hallucinatory yet socially impactful narrative that projects existential fears onto “gender ideology.” Duda’s 2020 re-election campaign solidified this approach. By stating that “LGBTIQ+ rights are more destructive than communism” and signing the Family Covenant (committing to ban same-sex marriage, halt gender education, and defend the “traditional family”), he turned gender into a symbolic threat to the Polish nation. This scene condenses various anxieties — about sovereignty, demographic decline, cultural change, and EU influence — into a single, emotionally charged enemy. Policies like the spread of “LGBTIQ-free zones” and the near-total abortion ban reinforce this projection, depicting feminists, sexual minorities, and their allies as threats to “the people” and the nation’s moral order. These measures stage a

psychosocial drama of fear and defence: citizens are encouraged to see themselves as besieged by foreign ideologies with Duda cast as the paternal protector.

In Brazil, during Jair Bolsonaro's presidency (2019-2022), anti-gender politics became a central part of the narrative used to stage authoritarian populism. Bolsonaro repeatedly invoked the threat of "gender ideology" as a corrupting influence endangering Brazilian children, families, and Christian morals. His administration promoted the "Escola Sem Partido" (School Without Party) initiative to ban discussions of gender and sexuality in classrooms, framing it as protection against indoctrination. LGBTIQ+ issues were removed from human rights policies, funding for films with queer themes was blocked and teachers faced intimidation for including feminism or sexuality in their curricula. These actions created a hallucinatory scenario in which widespread social fears—such as rising crime, economic instability, and the decline of patriarchal authority—were projected onto feminists, LGBTIQ+ communities and progressive educators. The "enemy within" was imagined as sexual deviance imposed by global elites, corrupting children's innocence and threatening the natural order. By depicting gender as a fundamental danger, Bolsonaro positioned himself as the paternal protector, defending Brazil's moral fabric against both domestic "subversives" and foreign conspiracies. This symbolic scene condensed crises of masculinity, religious conservatism and nationalism into a single struggle, enabling Bolsonaro to incite paranoia and justify authoritarian actions.

In Turkey, Recep Tayyip Erdoğan (2003-) and his political party AKP have increasingly used anti-gender politics as a tool to strengthen authoritarian populism. A pivotal moment was when they withdrew from the Istanbul Convention in 2021, the Council of Europe treaty aimed at preventing violence against women. Erdoğan justified this by portraying the Convention as an alien, Western imposition that threatened "Turkish family values." This act portrayed gender equality itself as a foreign ideology corrupting the nation. Beyond the treaty withdrawal, Erdoğan has consistently

amplified social anxieties related to gender. He has stated that “women and men are not equal by nature,” warned against “Western feminism,” and stressed that a woman’s primary role is motherhood. LGBTIQ+ rights have been portrayed as “perversions,” and Pride marches have been repeatedly banned and violently suppressed. These narratives portray feminists and sexual minorities as enemies, undermining the moral and demographic stability of the nation. By attributing broad crises—economic uncertainty, political division, declining birth rates—to the concept of “gender ideology,” Erdoğan positions himself as the paternal protector who reestablishes order through patriarchal authority and religious morality. Under Erdoğan’s leadership and with support from the Ministry of Family and Social Services, 2025 has been declared the “Year of the Family” to “protect and strengthen the family institution” against “harmful influences.” In this regard, “the increasing visibility of genderless ideologies and LGBTIQ+ narratives” was identified as a global risk to the family structure.

In the USA, Donald Trump’s presidency (2017-2021; 2025-) weaponised anti-gender politics by creating a false narrative where feminism, LGBTIQ+ rights and reproductive freedoms were portrayed as threats to “real America.” Trump repeatedly cast himself as the defender of the “forgotten man” and the “traditional family” against coastal liberal elites, feminists and “gender ideology.” This narrative simplified fears about demographic change, economic insecurity and cultural diversity into a distorted fight over sex, gender, and sexuality. Concrete policies underscored this narrative: Trump reinstated and expanded the Global Gag Rule, cutting U.S. funding to international organisations that supported abortion rights. His administration worked to weaken Title IX protections for victims of sexual assault on campuses and sought to define sex narrowly as biological and binary, effectively erasing trans rights in federal law. At the UN, U.S. diplomats under Trump pushed back against references to “sexual and reproductive health and rights,” aligning with conservative and religious states in global forums. Rhetorically, Trump demonised feminists as “nasty women,” mocked

survivors of sexual violence, and depicted trans people as threats to women and children in bathrooms, echoing the moral panic tactics used by other populist leaders. These actions attributed broad cultural and economic insecurities to “gender” and “sexual ideology,” portraying them as existential threats to American values. This dramatisation justifies regressive policies, energises conservative evangelical and right-wing supporters, and portrays Trump as the paternal protector of the nation.

Since the early 2010s, Vladimir Putin has made anti-gender politics a key part of his authoritarian consolidation, framing it as a scene where “gender ideology” represents the corruption of Western liberalism and the threat to Russian civilisation. The 2013 “gay propaganda law” banning the promotion of “non-traditional sexual relations” to minors marked a turning point: LGBTIQ+ people were portrayed as dangerous outsiders undermining children, family and national survival. In 2023, the Russian Supreme Court took it further by banning the LGBTIQ+ movement entirely, labelling it “extremist,” which was followed by police raids on queer spaces in Moscow. Putin repeatedly characterises feminism and LGBTIQ+ rights as Western imports that clash with Russian tradition and Orthodoxy. His rhetoric links various crises — such as declining demographics, economic instability, and geopolitical tensions — to “gender ideology,” framing it as a foreign plot aimed at weakening Russia. In his geopolitical rhetoric, Russia is depicted as the global guardian of “traditional values” against a decadent West, exporting anti-gender ideology as part of its soft power play.

Andrej Babiš, Prime Minister of the Czech Republic (2017-2021), represented an oligarchic style of authoritarian populism that employed anti-gender rhetoric to appeal to conservative voters and counter liberal opposition. Although less obvious than in Poland or Hungary, his government reflected regional patterns of framing “gender ideology” as a symbolic threat. In public debates about the Istanbul Convention, Babiš and his allies argued that ratifying the treaty would force an alien “gender theory” on Czech society, undermine traditional family values and weaken sovereignty. The

government delayed ratification under pressure from right-wing populist forces, rendering the treaty a proxy for concerns about EU interference and cultural liberalism. Babiš's movement ANO 2011 also exploited moral panics surrounding sex education and LGBTIQ+ rights. Right-wing media outlets and allied politicians portrayed debates over same-sex marriage and gender equality as “unnatural” or “foreign impositions.” This created a phantasmatic scene where diffuse anxieties—about corruption scandals, inequality, EU dependency, and political instability—were projected onto “gender ideology” as the enemy of the Czech family and nation. The surreal narrative of gender as an existential threat offered cultural cover for his power growth, normalising illiberal tendencies under the pretence of defending tradition. As with other cases, the patriarchal stability that Babiš's discourse evoked had never truly existed, but presenting it as a lost order legitimised resistance to progressive reforms and aligned Czechia with the broader Central European anti-gender wave led by Orbán and Kaczyński.

Since returning to power in 2010, Viktor Orbán has positioned Hungary as the center of anti-gender politics in Europe. Orbán depicts the EU, liberal elites, feminists, and LGBTIQ+ activists as corrupting influences that threaten the Hungarian family, nation and Christian civilisation. Concrete policies illustrate this dramatisation: In 2018, Hungary banned gender studies programs in universities, framing them as “ideology, not science.” In 2020, the government amended the constitution to define family strictly as based on marriage between a man and a woman, and limited adoption to heterosexual couples. In 2021, the Hungarian Parliament passed a law banning the “promotion” of homosexuality and gender transition to minors, linking LGBTIQ+ rights to pedophilia — a classic moral panic tactic. Orbán's government uses slogans like “Stop Brussels” and “Hungary must protect its children” to blend anti-gender politics with nationalist and anti-EU campaigns. In these efforts, multiple crises such as economic instability, rural inequality, and migration fears come together into a single symbolic battle against “gender.” By depicting gender as a life-threatening issue, Orbán

positions himself as the paternal guardian of the nation, justifying authoritarian policies and illiberal democracy.

Drawing on examples from around the world, here is evidence that anti-gender politics has played a crucial role in bringing authoritarian populists to power, solidifying their rule, and becoming an increasingly urgent global threat. These cases also demonstrate how anti-gender mobilisations have shifted from local cultural conflicts to an increasingly urgent global threat to democracy, human rights and gender equality. Since anti-gender discourses consistently portray the EU values and institutions as primary targets, the EU has a special responsibility to respond. It must therefore craft strong and clear policies that not only defend its core commitments to democracy, human rights, and gender equality, but also actively combat the spread of anti-gender mobilisations inside and outside its borders.

The EU has significantly shaped gender equality policies by guiding member states, organising its structure and acting globally. In 2019, the EU launched its own Action Plan on Women, Peace, and Security, aligning with United Nations Security Council Resolution 1325. This plan aims to mainstream gender equality and women's rights in the EU's external efforts, especially in conflict prevention, crisis management, and peacebuilding. The EU's Action Plan mainly targets prevention, protection, relief and recovery. Its goals include increasing gender expertise and focal points within EU institutions and Common Security and Defence Policy (CSDP) missions, raising funding for gender-sensitive programs, forming partnerships with civil society and women's organisations in conflict zones, and tracking progress through indicators, reports, and evaluations. The EU's Action Plan on Women, Peace, and Security complements its Gender Action Plan (GAP) III, which sets targets for including gender perspectives in external actions. In this way, the 2019 EU Action Plan strives to do more than just words by setting concrete goals for women's participation, protection, and empowerment, though it still faces challenges. Critics argue that the EU Action Plan is sometimes absorbed into traditional security agendas, like

counterterrorism and migration control, which can weaken its ability to promote real gender equality. Although the plan emphasises consultation, feminist and grassroots groups often say that engagement remains superficial rather than meaningful. Additionally, the EU's credibility is questioned because, even though it promotes WPS internationally, some member states (e.g., Poland, Hungary) pursue anti-gender policies at home, undermining consistency.

In this context, the EU must reassess and strengthen its normative power as a global promoter of democracy, peace and human rights. To address the anti-gender challenge, I propose three pillars, drawing from over a decade of feminist foreign policy experience: (1) Recognising the anti-gender movement as a global threat to human security, especially for women and LGBTIQ+ individuals, and prioritising efforts to combat this threat; (2) Supporting social movements, particularly those of women and LGBTIQ+ individuals, while creating avenues for their participation in foreign policy decision-making; (3) Adopting an intersectional approach to peacebuilding that addresses all forms of injustice and inequality by considering how gender intersects with other forms of inequality.

First, recognising anti-gender politics as a transnational threat to human security—especially for women and LGBTIQ+ communities—and including this analysis in peacebuilding and foreign policy strategies is urgent. It should be treated with the same urgency as other forms of extremism that threaten peace and democracy. Instead of seeing anti-gender movements as isolated or cultural phenomena, the EU must view them as part of a coordinated, ideologically driven challenge to liberal democratic values. Addressing this threat requires the EU to incorporate a gendered analysis into its security, foreign policy, and peacebuilding efforts. It is, of course, inaccurate to say that the EU has taken no action on this matter so far. The EU has recognised the rise of anti-gender rhetoric in its internal and external communications. The European Parliament has condemned anti-gender campaigns, especially in Poland and Hungary. Additionally, the 2020-2025 EU Gender Equality

Strategy explicitly links gender equality to democracy and the rule of law (European Commission, 2020). The Strategy commits to challenging gender stereotypes and countering sexist hate speech as key priorities. It emphasises intersectionality as a core principle for implementation, recognising that discrimination is complex and intersects with factors like gender, race and sexuality. The Strategy aims to incorporate a gender perspective across all EU policy areas. It also acknowledges that progress is neither guaranteed nor irreversible, suggesting that advances in gender equality encounter resistance. However, it does not explicitly reference “anti-gender politics,” “gender ideology,” or “backlash movements” as threats. It lacks a comprehensive framework within the Strategy to monitor or oppose anti-gender discourse or organised efforts against gender equality. It does not include language about defending rights from ideological or cultural attacks; the focus is more on structural inequalities, stereotypes, and gender balance. Because the Strategy does not directly address anti-gender movements, it may underestimate or lack adequate mechanisms to counter political backlash. Recognising that equality is not guaranteed forever is a weak acknowledgement of potential resistance, but it does not serve as a clear defence stance. Petra Debusscher (2023) argues that, although the Strategy makes progress, its legislative initiatives are modest, and many proposals have been stagnant for a long time, showing limited ambition in fighting resistance. As the European Commission faces increasing pressure from far-right and anti-gender groups in some Member States, the absence of stronger defensive language could leave the Strategy vulnerable.

On the other hand, the 2025 “Roadmap for Women’s Rights” explicitly signals the concern of the European Commission about backlash and contestation of gender equality (European Parliament, 2025). The Roadmap recognises “worrying trends that challenge existing gender equality and promote a sexist political discourse.” It partly aims to counteract political movements that oppose or try to roll back gender equality policies. It includes a declaration of principles for a gender-equal society, reaffirming the EU’s

commitment to protecting women's rights and urging all EU institutions to adhere to it. The Roadmap is designed to guide the next Gender Equality Strategy (post-2025) and is explicitly connected to addressing opposition and setbacks in gender equality. While the Roadmap extends beyond the 2020-2025 Strategy by acknowledging backlash and contestation, its language remains cautious — it does not employ strong, confrontational framing (e.g., labelling anti-gender politics as an existential threat). It falls short of proposing new measures or legislative initiatives; instead, it primarily serves as a guiding document, highlighting the challenge. Some civil society voices argue that it does not go far enough. For example, the End FGM European Network says the Roadmap “missed the opportunity to be stronger and more uncompromising” in its stance against anti-rights movements. The 2025 Roadmap is a significant symbolic step, but without stronger tools and an explicit acknowledgement of the anti-gender threat, it risks remaining largely symbolic. To be effective, the EU must shift from symbolic reaffirmation to establishing an institutionalised defence of gender equality as part of its larger fight against authoritarian populism.

To strengthen its external actions, the EU should also reevaluate how its Women, Peace, and Security (WPS) agenda and emerging FFP frameworks address the rise of anti-gender politics beyond Europe. Butler's concept of the phantasmatic scene shows that authoritarian regimes weaponise “gender ideology” as a hallucinated threat, justifying repression domestically and exporting illiberal values abroad. If the EU only promotes technical gender mainstreaming, it risks being dismissed as part of the “ideology” these actors demonise. Instead, the EU must frame WPS and FFP as tools to challenge anti-gender illusions — by emphasising how gender equality promotes real security, empowering civil society movements that oppose authoritarian narratives, and ensuring that intersectionality is genuinely integrated into peacebuilding and diplomacy. In doing so, the EU could reposition itself as a global actor that not only advances gender equality but also actively resists authoritarian uses of gender as a weapon of fear. Another problem is that the

EU often views anti-gender politics as a domestic issue within member states rather than a coordinated transnational threat. It is rarely integrated into foreign and security policies. Instead, the EU could officially recognise anti-gender mobilisations as a human rights concern in its foreign and development strategies and include counter-disinformation and strategic communication tools to combat anti-gender narratives.

In fact, recognising “gendered disinformation” as a foreign policy issue is essential for a comprehensive understanding of digital disinformation as a security threat (Hedling, 2024, 137). Digital disinformation campaigns, often spread by autocratic regimes or groups, weaponise sexist narratives within domestic debates as part of a foreign policy strategy (Hedling, 2024:138). Hedling (2024, 139) stresses that gendered disinformation is not marginal; it is increasingly acknowledged as a widespread, deliberate, and cross-cutting tactic used by actors aiming to weaken the core of democratic systems. In an era marked by populism and a regressive attitude toward gender equality in many nations, Hedling (2024) explores how gendered disinformation functions as a “divide and conquer” tactic. This approach frequently employs divisive gender narratives to divert attention from other pressing social and political issues, manipulating language and imagery to reinforce biases. This further solidifies discriminatory practices and deepens societal divisions (Hedling, 2024, 142). Over the past decade, rising concerns about this issue have prompted countries like the USA and the UK, along with organisations such as the UN and the EU, to issue statements and reports acknowledging gendered disinformation as a significant foreign policy issue and a global threat to democracy. Nonetheless, more gender-informed analyses are necessary to effectively address this form of influence and the vulnerabilities that sustain it (Hedling, 2024, 139).

Secondly, supporting grassroots movements—especially those led by women and LGBTIQ+ individuals—by creating meaningful ways for them to participate in foreign policy and peacebuilding efforts is crucial for the EU to reevaluate its gender policies amid anti-gender politics. These groups are

often the first to confront authoritarian and exclusionary politics. To effectively counter anti-gender politics, the EU must do more than just rhetorical support; it must actively fund and legitimise feminist and LGBTIQ+ social movements as drivers of democratic renewal. This includes providing sustainable funding, legal protections, and transnational networks for grassroots activism. Equally important is developing institutional mechanisms—such as advisory councils, consultative platforms, and co-decision processes—that enable these movements to influence foreign policy agendas. Their inclusion boosts democratic legitimacy, improves policy responsiveness, and strengthens the EU's role as a defender of pluralism and rights-based governance.

The EU has taken significant steps on this issue to date, supporting civil society through programs such as the European Instrument for Democracy and Human Rights (EIDHR), the NDICI – Global Europe, and the Citizens, Equality, Rights, and Values (CERV) program. EIDHR was established in 2006, with a focus on supporting human rights defenders and pro-democracy civil society groups, particularly in repressive regimes where EU delegations could operate independently of governments. It funded projects on gender equality, combating violence against women, and LGBTIQ+ rights. It was replaced by NDICI in 2021, but its legacy continues as part of NDICI's thematic programmes. Launched in 2021 for the 2021-2027 budget cycle, NDICI–Global Europe is the EU's primary external action funding tool, with a €79.5 billion budget. It allocates funds explicitly to human rights defenders, gender equality, LGBTIQ+ rights, and women's organisations, both inside and outside conflict zones. It includes the “Thematic Programme on Human Rights and Democracy,” which directly funds NGOs, grassroots feminist movements, and civil society groups without requiring them to go through national governments (a key feature in authoritarian contexts). CERV is another vital instrument. It is the EU's most comprehensive internal funding program for rights and democracy, running from 2021 to 2027 with a budget of €1.55 billion. It supports civil society organisations within the EU,

especially those working on gender equality, anti-discrimination, LGBTIQ+ rights, the rule of law and civic participation. This support covers initiatives combating anti-gender movements, projects advancing gender equality in member states and networks defending LGBTIQ+ rights. The European Endowment for Democracy (EED), a flexible funding tool that supports pro-democracy activists, including women's and LGBTIQ+ groups, especially in the EU neighbourhood, will also be considered, along with other relevant instruments. Through these instruments, the EU has built one of the largest global funding ecosystems for civil society with a strong emphasis on gender equality and LGBTIQ+ rights.

These instruments are crucial for countering anti-gender movements because they enable direct support to grassroots actors, bypass hostile governments, and reinforce the EU's external role as a defender of democracy and human rights. However, funding often remains short-term, bureaucratic and difficult for grassroots movements—especially in the Global South—to access. Application and reporting procedures tend to be too demanding for small grassroots groups. This benefits large and professionalised NGOs and can marginalise the very movements—such as feminist collectives, queer activists, and local women's groups—that are most impacted by anti-gender backlash. In authoritarian contexts, EU funding can expose recipients to risks, like state harassment, stigmatisation or of being labelled “foreign agents” (e.g., Russia, Hungary). Anti-gender groups exploit this by portraying EU-funded NGOs as illegitimate elites imposed from abroad. Additionally, while CERV aims to promote equality within the EU, in countries like Poland and Hungary, governments have restricted access or delegitimised EU-funded organisations. This highlights the gap between EU-level commitments and national realities and limits the success of the LGBTIQ Equality Strategy 2020-2025, which seeks to support LGBTIQ+ organisations and protect their rights.

In other words, the EU's civil society funding programs are vital for countering anti-gender mobilisations, but they are also vulnerable. Without

reforms to simplify access, guarantee long-term sustainability, and protect partners from authoritarian retaliation, these tools risk merely reproducing the very fragility they aim to address. If the EU wants to be a credible defender of democracy and gender equality, it must do more than just fund projects. It also needs to invest in movement resilience, ensuring that feminist and LGBTIQ+ actors can withstand authoritarian and anti-gender pushback over the long term. Easing bureaucratic barriers in calls for proposals and reporting requirements—especially for small grassroots organisations—and creating user-friendly application systems, along with providing technical support like translation and training for local actors who may lack professional grant-writing skills, will facilitate access. Moving beyond short-term project cycles by offering core, multi-year institutional funding to feminist, LGBTIQ+, and human rights organisations, and expanding micro-grant programs with rapid disbursement for urgent activism (e.g., responding to sudden state crackdowns) will foster more flexible and sustainable funding. Another necessity is connecting civil society support with broader rule-of-law conditionality tools—such as suspending EU funds to member states that suppress feminist or LGBTIQ+ NGOs—and ensuring that CERV funds are not blocked or undermined by hostile governments in countries like Poland or Hungary, thereby linking funding to rule-of-law mechanisms.

Furthermore, although the EU has created some mechanisms to amplify the voices of women's and LGBTIQ+ movements in foreign policy decisions, these are still limited in scope and impact. What is needed is a systematic institutionalisation of participatory processes. For example, the EU could establish permanent consultative forums or foreign policy advisory groups that formally include feminist and LGBTIQ+ civil society organisations in agenda-setting, policy design, and evaluation. Such structures would ensure that equality concerns are not seen as external additions but as core elements in shaping EU external action. Additionally, the EU should supplement participation with practical support mechanisms. This could include rapid-access funding options for feminist and LGBTIQ+ advocates working in

hostile or authoritarian settings, enabling them to respond quickly to crises. Equally, capacity-building programs that enhance the advocacy, security, and organisational skills of grassroots actors, enabling them to engage effectively with EU institutions and international organisations, are crucial. By connecting participatory structures with sustainable support and protection measures, the EU would go beyond symbolic inclusion to foster an environment where civil society can play a vital role in shaping FFP.

Thirdly, adopting an intersectional approach to peacebuilding that recognises overlapping systems of oppression—such as gender, race, class, migration status, or sexuality—is essential for creating a genuinely transformative gender policy, especially in response to the anti-gender backlash. As Thompson and Clement (2020, 5) note, a common mistake is for institutions to claim their policies as feminist by focusing solely on girls and women or by considering a quota of women as enough, without addressing intersectionality. A sustainable and just peace cannot be achieved without understanding how gender-based inequalities are compounded by race, class, ethnicity, migration status, disability, and other axes of marginalisation. The Gender Action Plan III (2021-2025) explicitly commits to intersectionality and gender mainstreaming across EU external actions. However, the plan pays limited attention to intersecting inequalities like race, class, sexuality, and disability, which reduces its overall inclusiveness. Some EU-funded peace and development projects also include gender-disaggregated data and participatory assessments. But intersectionality is often a rhetorical pledge; in practice, implementation varies across EU institutions and delegations. Data collection remains limited, especially regarding race, migration status and other intersecting forms of inequality.

Maes and Debusscher (2024) examine how the idea of intersectionality has been portrayed in the EU three Gender Action Plans (GAP I: 2010-2015; GAP II: 2016-2020; GAP III: 2021-2025) for external relations. In GAP I, there is no mention of intersectionality; women are viewed as a uniform group. GAP II introduces intersectionality, but mainly as an additive concept (gender +

other categories = additional discrimination). The focus remains on women as a single group. GAP III shows a significant change, explicitly identifying intersectionality as a core principle. The plan recognises multiple disadvantages (e.g., race, age, disability, migration, and LGBTIQ+ status), emphasises inclusivity, and increases consultation with diverse civil society groups. Still, it often treats intersectionality more as an individual identity issue than a structural one. It overlooks reflection on the EU's own role in perpetuating inequalities (e.g., through trade or migration policies). Maes and Debusscher (2024) argue that the EU has developed a more comprehensive and explicit approach to intersectionality in its external gender policy, especially under GAP III. This marks a significant advance, giving advocates more influence and demonstrating the EU's goal to be a progressive global leader in gender issues. However, the EU's framing of intersectionality remains limited and lacks political nuance: it tends to highlight individual disadvantages rather than examining broader systems of power (e.g., sexism, capitalism, colonialism); it often overlooks privilege and relational dynamics, focusing only on marginalised groups without addressing dominant positions; and it lacks critical self-reflection regarding the EU's own policies and its role in maintaining inequalities worldwide.

In short, while the EU has made progress in addressing various inequalities, these are often implicit and poorly articulated. A more systematic, explicit and transformative approach to intersectionality is necessary to improve the inclusiveness and effectiveness of EU equality policies. The EU should adopt intersectionality not only as an analytical tool but also as a guiding principle in shaping and executing its foreign, development and security strategies. This includes data disaggregation, participatory methods and policy coherence across different fields. By tackling the root causes of structural inequality in a unified way, the EU can promote a more inclusive vision of peace that resonates with diverse communities and strengthens democratic solidarity.

5. Conclusions

This paper has demonstrated that anti-gender politics are not just fringe culture wars but are central tools of authoritarian populism. As Butler's concept of the phantasmatic scene shows, authoritarian leaders in Poland, Brazil, Turkey, the USA, Russia, Czechia, and Hungary depict "gender ideology" as a hallucinated threat, transforming broad social anxieties into existential battles over family, nation and sovereignty. In doing so, they harness fear and nostalgia to justify illiberal governance. The power of this dynamic lies precisely in its fantasy nature: the patriarchal order that authoritarian populists vow to restore never fully existed, yet its envisioned revival fuels the emotional core of their politics.

The EU is not just an external observer of this phenomenon but a main target. Anti-gender movements across Europe regularly portray EU values and institutions as hostile impositions, depicting Brussels as the face of "gender ideology." The EU therefore has a special responsibility: to defend its core commitments to democracy, human rights and gender equality against organised backlash. While the 2020-2025 Gender Equality Strategy and the 2025 Roadmap for Women's Rights reaffirm these values, they risk staying symbolic without stronger tools. If the EU does not directly challenge the illusionary logic, authoritarian populists will keep setting the terms of the debate.

To meet this challenge, the EU must move beyond rhetorical reaffirmation to institutionalised defence. This involves explicitly identifying anti-gender movements as coordinated illiberal actors; connecting gender equality to rule of law mechanisms and funding conditions; and providing direct, ongoing support to feminist and LGBTQ+ organisations. It also requires developing counter-narratives that reveal the falsehood of "gender ideology" myths, while presenting inclusive and intersectional visions of democracy and security.

Externally, the EU should adapt its Women, Peace, and Security agenda and emerging FFP frameworks to meet this challenge. Instead of treating gender equality as just a technical add-on, the EU must view it as a strategic tool to counter authoritarian illusions that exploit gender to spread fear. By empowering grassroots actors and integrating intersectionality into peacebuilding and diplomacy, the EU can strengthen its role as a global leader in promoting equality and democracy.

In short, anti-gender politics are an increasing global threat, and the EU's credibility depends on its ability to recognise them. By challenging the illusory scene that fuels authoritarian populism—both inside and outside its borders—the EU can turn gender equality from a fragile value into a strong pillar of democratic resilience. The EU has played a vital role in shaping gender equality policies, guiding member states, influencing its organisational structure, and acting internationally. However, unless it uses this institutional power to create inclusive and effective ways to fight anti-gender movements, progress will not meet expectations. Its efforts in promoting gender equality and human rights have been significant, but its ability to resist and counter anti-gender mobilisation remains underdeveloped.

In conclusion, the rise of anti-gender politics and authoritarian populism presents a major challenge to the EU's identity and core values. As the global political landscape changes, the EU must reaffirm its commitment to gender justice, not only as a matter of rights but as the foundation of sustainable peace and democracy. By adopting a reimagined FFP that is intersectional, inclusive and resilient, the EU can lead in safeguarding human security and democratic principles for all.

References

- Aggestam K. and Bergman-Rosamond A. (2016). Swedish Feminist Foreign Policy in the Making: Ethics, Politics, and Gender, in *Ethics & International Affairs*, Vol. 30, No. 3.
- Aggestam K., Rosamond A. B. and Hedling E. (2024). *The Politics of Feminist Foreign Policy and Digital Diplomacy* (Springer Nature).
- Akgemci, E. (2022). Authoritarian Populism as a Response to Crisis: The Case of Brazil, in *International Relations*, Vol. 19, No. 74.
- Bauman Z. (2017). *Retrotopia* (Polity Press).
- Butler J. (1990). *Gender Trouble* (Routledge).
- Butler J. (2021). Why is the Idea of ‘Gender’ Provoking Backlash the World Over?”, *The Guardian*, October 23.
- Butler J. (2025). *Who’s Afraid of Gender?* (Random House).
- De Beauvoir S. (1949). *Le Deuxième Sexe* (Éditions Gallimard).
- Debusscher P. (2023). The EU Gender Equality Strategy 2020-2025: The Beginning of a New Season?, in *Social Policy in the European Union: State of Play 2022: Policymaking in a Permacrisis* (ETUI; OSE).
- Dempsey D. M. (2020). *Interrogating Pope Francis: On Gender Theory and Ideological Colonization* (University of California, Riverside).
- Deslandes A. (2020). Checking in on Mexico’s Feminist Foreign Policy, in *Foreign Policy*, No. 30.
- Dietze G. and Roth J. (2020). Right-Wing Populism and Gender: A Preliminary Cartography of an Emergent Research Field, in G. Dietze and J. Roth (eds.), *Right-Wing Populism and Gender: European Perspectives and Beyond* (Transcript).
- European Commission (2020). *A Union of Equality: Gender Equality Strategy 2020-2025*, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC015-2>
- European Parliament (2025). *Roadmap for Women’s Rights Next steps for EU Action on Gender Equality*,

- https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/769542/-EPRS_BRI%282025%29769542_EN.pdf?utm_source=chatgpt.com
- Fassin É. (2020). Anti-Gender Campaigns, Populism, and Neoliberalism in Europe and Latin America, in *LASA Forum*, Vol. 51, No. 2.
- Foster S. and Markham S. A. (2024). *Feminist Foreign Policy in Theory and in Practice: An Introduction* (Routledge).
- German Federal Foreign Office (2023). *Shaping Feminist Foreign Policy – Federal Foreign Office Guidelines*,
https://www.shapingfeministforeignpolicy.org/papers/Guidelines_Feminist_Foreign_Policy.pdf.
- Gokay B., Xypolia I. and Mandelbaum M. (2017). The Rise of ‘Authoritarian Populism’ in the 21st Century: From Erdoğan's Turkey to Trump's America, in *Journal of Global Faultlines*, Vol. 4, No. 1.
- Government of Canada (2018). *Canada's Feminist International Assistance Policy*, https://www.international.gc.ca/world-monde/issues_developpement-enjeux_developpement-ent/priorities-priorites/policy-politique.aspx?lang=eng.
- Government of Chile (2022). *Feminist Foreign Policy*, <https://drive.google.com/file/d/1gfJREM-5CbejM5F2ADAprE3NW79vCVIW/view>.
- Government of France (2025). *France's International Strategy for a Feminist Foreign Policy (2025-2030)*, https://www.diplomatie.gouv.fr/en/french-foreign-policy/feminist-diplomacy/france-s-international-strategy-for-a-feminist-foreign-policy-2025-2030/?utm_source=chatgpt.com.
- Government of Luxembourg (2025). *Second National Plan “Women, Peace and Security” 2025-2030*, <https://mae.gouvernement.lu/dam-assets/directions/d1/pan-femm-es-paix-securite/second-national-action-plan-wps-en.pdf>.
- Government of Mexico (2020). *Mexico Adopts Feminist Foreign Policy*. <https://www.gob.mx/sre/prensa/mexico-adopts-feminist-foreign-policy?idiom=en>.

Government of the Netherlands (2022). *Feminist Foreign Policy Explained*, <https://www.government.nl/latest/news/2022/11/18/feminist-foreign-policy-netherlands>.

Government of Spain (2021). *Spain's Feminist Foreign Policy: Promoting Gender Equality in Spain's External Action* , https://www.exteriores.gob.es/es/Servicios-AlCiudadano/PublicacionesOficiales/2021_02_POLITICA%20EXTERIOR%20FEMINISTA_ENG.pdf?utm_source=chatgpt.com.

Government of Sweden (2015). *Statement of Foreign Policy 2015*, <https://www.government.se/speeches/2015/02/statement-of-foreign-policy-2015/>.

Government of Sweden (2020). *The Government's Statement of Foreign Policy 2020*, <https://www.government.se/speeches/2020/02/2020-statement-of-foreign-policy/>.

Graff A. and Korolczuk E. (2022). *Anti-Gender Politics in the Populist Moment* (Taylor & Francis).

Hall S. et al. (1978). *Policing the Crisis: Mugging, the State and Law and Order* (Palgrave Macmillan).

Hall S. (1988). *The Hard Road to Renewal: Thatcherism and the Crisis of the Left* (Verso).

Hauschild A. and Stamm L. (2024). From Feminist Questions towards Feminist Processes: Strengthening Germany's Feminist Foreign Policy, in *European Journal of Politics and Gender*, Vol.7, No. 2.

Hedling E. (2024). Gendered Disinformation, in K. Aggestam and J. True (eds.) *Feminist Foreign Policy Analysis: A New Subfield* (Bristol University Press).

Kandiyoti D. (2022). How Did Gender Move to the Center of Democratic Struggles?, in *Seminar, Kapuscinski Lectures by Columbia Global Center*.

Kourou N. S. (2022). Right-Wing Populism and Anti-Gender Movements: The Same Coin with Different Faces, in *Global Political Trends Center (GPoT)*.

- Kuhar R. (2023). The Rise and Success of the Anti-Gender Movement in Europe and Beyond, in *Progressive Yearbook*.
- Kuhar R. and Paternotte D. (2017). “Gender Ideology” in Movement: Introduction, in R. Kuhar and Paternotte D. (eds.), *Anti-Gender Campaigns in Europe: Mobilizing Against Equality* (Bloomsbury Publishing PLC).
- Maes E. L. and Debusscher P. (2024). The EU as a Global Gender Actor: Tracing Intersectionality in the European Gender Action Plans for External Relations 2010–2025, in *Social Politics: International Studies in Gender, State & Society*, Vol. 31, No. 1.
- Mehring K., Off G. and Wojnicka K. (2025). The Populist Radical Right, the Gender Gap and Protective Masculinity Across European Countries, in *European Journal of Politics and Gender*, Vol. 20, No. 20.
- Morelock J. and Ziotti Narita F. (2021). A Dialectical Constellation of Authoritarian Populism in the United States and Brazil, in J. Morelock (ed.), *How to Critique Authoritarian Populism* (Brill).
- Norris P. and Inglehart R. (2019) *Cultural Backlash: Trump, Brexit, and Authoritarian Populism* (Cambridge University Press).
- Novović G. (2024). Fit for feminism? Examining Policy Capacity for Canada’s Feminist Foreign Policy, in *Canadian Foreign Policy Journal*, Vol. 30, No. 3.
- Oakley A. (1972). *Sex, Gender and Society* (Temple Smith).
- Parker E. (2022). Heart in the Right Place: Thatcherism and Love in Jeanette Winterson’s *The Passion*, in *Contemporary Women's Writing*, Vol. 16, No. 3.
- Pierobon C. (2024). Shaping German Feminist Foreign Policy in Times of Conflict in Ukraine, in A. Mihr and C. Pierobon (Eds.), *Polarization, Shifting Borders and Liquid Governance* (Springer).
- Rich A. (1980). Compulsory Heterosexuality and Lesbian Existence, in *Signs*, Vol. 5, No. 4.
- Rubin G. (1975). The Traffic in Women: Notes on the “Political Economy” of Sex, in R. R. Reiter (ed.), *Toward an Anthropology of Women* (Monthly Review Press).

- Schleusener S. (2020). "You're Fired!" Retrotopian Desire and Right-Wing Class Politics, in G. Dietze and J. Roth (eds.) *Right-Wing Populism and Gender: European Perspectives and Beyond* (Transcript).
- Sedgwick E. K. (1985). *Between Men: English Literature and Male Homosocial Desire* (Columbia University Press).
- Scott J. (1986). Gender: A Useful Category of Historical Analysis, in *American Historical Review*, No. 91.
- Scott S.V. and Bloomfield A. (2017). Norm Entrepreneurs and Antipreneurs: Chalk and Cheese, or Two Faces of the Same Coin?, in A. Bloomfield and S.V. Scott (eds.) *Norm Antipreneurs and the Politics of Resistance to Global Normative Change* (Routledge).
- Thompson L. and Clement R. (2020). Definiendo La Política Exterior Feminista, in *International Center for Research on Women*.
- Thompson L., Ahmed S. and Khokhar T. (2021). Defining Feminist Foreign Policy: A 2021 Update, in *International Center for Research on Women*.
- Thomson J. and Whiting S. (2022). Women, Peace and Security National Action Plans in Anti-Gender Governments: The Cases of Brazil and Poland, in *European Journal of International Security*, Vol. 7, No. 4.
- Thomson J. (2024). Norms, in K. Aggestam and J. True (eds.) *Feminist Foreign Policy Analysis: A New Subfield* (Bristol University Press).
- Thomson J. and Wehner L. (2025). The Adoption of Feminist Foreign Policy: The Cases of Chile and Sweden. *Politics & Gender*.
- Towns A., Jezierska K. and Bjarnegård E. (2024). Can a Feminist Foreign Policy Be Undone? Reflections from Sweden, in *International Affairs*, Vol. 100, No. 3.

ATHENA

CRITICAL INQUIRIES IN LAW, PHILOSOPHY AND GLOBALIZATION

[HTTPS://ATHENA.UNIBO.IT/](https://athena.unibo.it/)

ATHENA@UNIBO.IT